Damm Distribución Integral

Business Analytics Project

Babak Barghi, Han Jia, Alexander Rutten

June 02, 2021

Contents

1 R Setup

The analysis is carried out in *R 4.0.2* and packages below are used.

```
library(tidyverse)
library(kableExtra)
library(RColorBrewer)
library(skimr)
library(scales)
library(tidymodels)
library(arules)
library(arulesViz)
library(waffle)
my_colors <- RColorBrewer::brewer.pal(6, "OrRd")[4:9]
```

2 Introduction

In this report we're going to review the data that DDI provided to us. This data is about their sales of food and drinks in a few regions in Barcelona. We have to review this data according to a few challenges that DDI gave us:

How many bars and restaurants only buy beer or soft drinks to DDI? How much wine or coffee can be sold in the designated areas? *For each unit of beer sold, how many units of soft drinks, wine and coffee are sold respectively?

By answering these challenges we will know more about the regions and how they can be significant for DDI. With this knowledge we will try to take on the last challenge:

Could you recommend something completely new about the DDI operations in these districts based on the data?

With the end result we will help DDI by better understanding their data and how this can help DDI in the future.

2.1 Damm Distribución Integral

Damm Distribución Integral (DDI) is a distribution company from Damm. Damm is a Spanish brewery founded in Barcelona in 1876. They're mainly known for their brands in beer like Estrella Damm, but have a lot of other brands that they sell and distribute across Spain and the world.

To distribute all their products they've their own distribution company, Damm Distribución Integral (DDI). DDI is made up of more than 25 distribution companies dedicated to the hospitality channel with a multi-product portfolio of more than 12.000 products and references. The company has over 15 years of experience and 53 specialists and sales representatives. With these 15 years of experience they're able to make 255 commercial routes across Spain to help 47.000 clients in major and smaller cities.

For our project we will focus on the area of Barcelona. In this area Distridam does the distribution of the products. In Barcelona they have around 9000 clients and more than 3000 products and references they sell to them. Distridam has 34 specialists and sales representatives and 65 Last-Mile delivery trucks to assist them.

The typical client for DDI will be a traditional barista with Basic Academic Training between 30-55 years old. This person is energetic, in neighborhood life and uses little digitalization.

2.2 Data Overview

DDI has provided us with their own dataset. This dataset contains all the necessary data of their clients in the needed areas of Barcelona. All the purchases are sorted by client. These clients are sorted by the areas that are needed for this project.

The clients, whose names are given, are divided in different establishments, such as restaurants and bars.

Per client we can exactly see which products they buy from DDI. These products can be divided by type of product, product code and product name. For all these products we can see how many they bought, how many liters/kilos or how much they cost for the client.

The quantities can be seen by year, the years 2019 and 2020, or by each month of those years.

```
#importing the data set
raw_data <- read_csv("raw-data2.csv",
                     col_types = cols(
                       `Total UMB` = col_character()))
```

```
#Changing Total UMB type to numeric
raw_data$`Total UMB` <- as.numeric(gsub(",", "", raw_data$`Total UMB`))
```

```
#Check the NAs
sum(is.na(raw_data))
```

```
## [1] 0
```

As we see there are no missing values in dataset, thus before starting with the analysis we would take a closer look at data frame using *glimpse* function.

```
#Close look
glimpse(raw_data)
```

```
## Rows: 94,581
## Columns: 85
## $ C.Postal                   <chr> "08001", "08001", "08001", "08001", "08001"~
## $ `Tipo establecimiento`     <chr> "ALIMENTACION TRADICIONAL", "ALIMENTACION T~
## $ `Cliente(dist)`            <chr> "BONA COMPRA - 29619", "BONA COMPRA - 29619~
## $ `Tipo de venta`            <chr> "Facturas", "Facturas", "Facturas", "Factur~
## $ `Línea de Negocio (dist)`  <chr> "BATIDOS", "BATIDOS", "AGUAS", "ALIMENTACIO~
## $ `Codigo Producto`          <chr> "40005", "30222", "20010", "85232", "30239"~
## $ Producto                   <chr> "CACAOLAT 1L SR 6U - 40005", "CACAOLAT MINI~
## $ `Total UMB`                <dbl> 0, 0, 143, 6, 1, 110, 14, 3, 9, 2, 1, 1, 2,~
## $ `Total UMB  ant.`          <dbl> 0, 0, 99, 0, 0, 0, 0, 0, 0, 14, 20, 13, 22,~
## $ `Total Litros Kilos`       <dbl> 0, 0, 1716, 0, 2, 871, 84, 18, 54, 12, 6, 6~
## $ `Total Litros Kilos  ant.` <dbl> 0, 0, 1188, 0, 0, 0, 0, 0, 0, 84, 120, 78, ~
## $ `Total Euros`              <dbl> 0.00, 0.00, 1081.25, 7.39, 6.61, 1045.00, 8~
## $ `Total Euros  ant.`        <dbl> 0.00, 0.00, 744.84, 0.00, 0.00, 0.00, 0.00,~
## $ `ene UMB`                  <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `ene UMB  ant.`            <chr> "-", "-", "0", "-", "-", "-", "-", "-", "-"~
## $ `ene Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `ene Litros Kilos  ant.`   <chr> "-", "-", "0", "-", "-", "-", "-", "-", "-"~
## $ `ene Euros`                <chr> "-", "-", "88.25", "-", "-", "-", "-", "-",~
```

```
## $ `ene Euros  ant.`          <chr> "-", "-", "0.00", "-", "-", "-", "-", "-", ~
## $ `feb UMB`                  <chr> "0", "0", "11", "-", "-", "-", "-", "-", "-~
## $ `feb UMB  ant.`            <chr> "0", "0", "0", "-", "-", "-", "-", "-", "-"~
## $ `feb Litros Kilos`         <chr> "0", "0", "132", "-", "-", "-", "-", "-", "~
## $ `feb Litros Kilos  ant.`   <chr> "0", "0", "0", "-", "-", "-", "-", "-", "-"~
## $ `feb Euros`                <chr> "0.00", "0.00", "82.75", "-", "-", "-", "-"~
## $ `feb Euros  ant.`          <chr> "0.00", "0.00", "0.00", "-", "-", "-", "-",~
## $ `mar UMB`                  <chr> "-", "-", "11", "-", "-", "-", "2", "-", "2~
## $ `mar UMB  ant.`            <chr> "-", "-", "0", "-", "-", "-", "0", "-", "0"~
## $ `mar Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "12", "-", ~
## $ `mar Litros Kilos  ant.`   <chr> "-", "-", "0", "-", "-", "-", "0", "-", "0"~
## $ `mar Euros`                <chr> "-", "-", "82.75", "-", "-", "-", "11.95", ~
## $ `mar Euros  ant.`          <chr> "-", "-", "0.00", "-", "-", "-", "0.00", "-~
## $ `abr UMB`                  <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-"~
## $ `abr UMB  ant.`            <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-"~
## $ `abr Litros Kilos`         <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-"~
## $ `abr Litros Kilos  ant.`   <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-"~
## $ `abr Euros`                <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-"~
## $ `abr Euros  ant.`          <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-"~
## $ `may UMB`                  <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `may UMB  ant.`            <chr> "-", "-", "0", "-", "-", "-", "-", "-", "-"~
## $ `may Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `may Litros Kilos  ant.`   <chr> "-", "-", "0", "-", "-", "-", "-", "-", "-"~
## $ `may Euros`                <chr> "-", "-", "82.75", "-", "-", "-", "-", "-",~
## $ `may Euros  ant.`          <chr> "-", "-", "0.00", "-", "-", "-", "-", "-", ~
## $ `jun UMB`                  <chr> "-", "-", "11", "6", "1", "45", "1", "1", "~
## $ `jun UMB  ant.`            <chr> "-", "-", "11", "0", "0", "0", "0", "0", "-~
## $ `jun Litros Kilos`         <chr> "-", "-", "132", "0", "2", "356", "6", "6",~
## $ `jun Litros Kilos  ant.`   <chr> "-", "-", "132", "0", "0", "0", "0", "0", "~
## $ `jun Euros`                <chr> "-", "-", "82.75", "7.39", "6.61", "427.50"~
## $ `jun Euros  ant.`          <chr> "-", "-", "82.76", "0.00", "0.00", "0.00", ~
## $ `jul UMB`                  <chr> "-", "-", "22", "-", "-", "65", "1", "1", "~
## $ `jul UMB  ant.`            <chr> "-", "-", "22", "-", "-", "0", "0", "0", "-~
## $ `jul Litros Kilos`         <chr> "-", "-", "264", "-", "-", "515", "6", "6",~
## $ `jul Litros Kilos  ant.`   <chr> "-", "-", "264", "-", "-", "0", "0", "0", "~
## $ `jul Euros`                <chr> "-", "-", "165.50", "-", "-", "617.50", "5.~
## $ `jul Euros  ant.`          <chr> "-", "-", "165.52", "-", "-", "0.00", "0.00~
## $ `ago UMB`                  <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `ago UMB  ant.`            <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `ago Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `ago Litros Kilos  ant.`   <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `ago Euros`                <chr> "-", "-", "82.75", "-", "-", "-", "-", "-",~
## $ `ago Euros  ant.`          <chr> "-", "-", "82.76", "-", "-", "-", "-", "-",~
## $ `sep UMB`                  <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `sep UMB  ant.`            <chr> "-", "-", "22", "-", "-", "-", "-", "-", "-~
## $ `sep Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `sep Litros Kilos  ant.`   <chr> "-", "-", "264", "-", "-", "-", "-", "-", "~
## $ `sep Euros`                <chr> "-", "-", "82.75", "-", "-", "-", "-", "-",~
## $ `sep Euros  ant.`          <chr> "-", "-", "165.52", "-", "-", "-", "-", "-"~
## $ `oct UMB`                  <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `oct UMB  ant.`            <chr> "-", "-", "11", "-", "-", "-", "-", "-", "-~
## $ `oct Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `oct Litros Kilos  ant.`   <chr> "-", "-", "132", "-", "-", "-", "-", "-", "~
## $ `oct Euros`                <chr> "-", "-", "82.75", "-", "-", "-", "-", "-",~
```

```
## $ `oct Euros  ant.`          <chr> "-", "-", "82.76", "-", "-", "-", "-", "-",~
## $ `nov UMB`                  <chr> "-", "-", "22", "-", "-", "-", "7", "1", "5~
## $ `nov UMB  ant.`            <chr> "-", "-", "11", "-", "-", "-", "0", "0", "0~
## $ `nov Litros Kilos`         <chr> "-", "-", "264", "-", "-", "-", "42", "6", ~
## $ `nov Litros Kilos  ant.`   <chr> "-", "-", "132", "-", "-", "-", "0", "0", "~
## $ `nov Euros`                <chr> "-", "-", "165.50", "-", "-", "-", "41.83",~
## $ `nov Euros  ant.`          <chr> "-", "-", "82.76", "-", "-", "-", "0.00", "~
## $ `dic UMB`                  <chr> "-", "-", "11", "-", "-", "-", "3", "-", "2~
## $ `dic UMB  ant.`            <chr> "-", "-", "11", "-", "-", "-", "0", "-", "0~
## $ `dic Litros Kilos`         <chr> "-", "-", "132", "-", "-", "-", "18", "-", ~
## $ `dic Litros Kilos  ant.`   <chr> "-", "-", "132", "-", "-", "-", "0", "-", "~
## $ `dic Euros`                <chr> "-", "-", "82.75", "-", "-", "-", "17.93", ~
## $ `dic Euros  ant.`          <chr> "-", "-", "82.76", "-", "-", "-", "0.00", "~
```

The dataset *raw_data* provides 94581 observations with 85 different variables.

2.3 Prepare Data

Before getting into the main analysis, we would apply some data manipulation to prepare the data frame for further instructions.

```
#change column names
ddi <-
  raw_data %>% rename(
  Postal_Code = `C.Postal`,
  Store_type = `Tipo establecimiento`,
  Store_name = `Cliente(dist)`,
  Sale_type = `Tipo de venta`,
  Product_line = `Línea de Negocio (dist)`,
  Product_code = `Codigo Producto`,
  Product_name = Producto,
  Total_units20 = `Total UMB`,
  Total_units19 = `Total UMB  ant.`,
  Total_amount20 = `Total Litros Kilos`,
  Total_amount19 = `Total Litros Kilos  ant.`,
  Total_revenue20 = `Total Euros`,
  Total_revenue19 = `Total Euros  ant.`
)
```

```
#Name of area
ddi$Postal_Code <- revalue(ddi$Postal_Code,
                           c("08001"="Gotic",
                             "08002"="Raval",
                             "08003"="City Center",
                             "08014"="Sants",
                             "08028"="Les Corts",
                             "08860"="Castelldefels",
                             "08830"="Sant Boi"))
```

```
## Warning: 'plyr' namespace cannot be unloaded:
##   namespace 'plyr' is imported by 'pROC' so cannot be unloaded
```

3 Challenges

3.1 Challenge 1

For this challenge we need to find out how many bars and restaurants only buy beer or soft drinks from DDI.

Which data did we use? We used the data about what kind of category the products are. We did this because we only needed to work with the categories beer and soft drinks. We also needed to use the bar/restaurant names. This way we could get all the bars and restaurants that only bought beer and soft drinks in the years 2019 and 2020. In order to make a proper assumption regarding all the product types, we had to set them in the different categories.

```
#categorize products
ddi <-
  ddi %>% mutate(Product_category = case_when(
  grepl("ENVASES", Product_line) ~ "Packaging",
  grepl("AGUA", Product_line) ~ "Water",
  grepl("CERVEZA", Product_line) ~ "Beer",
  grepl("GASEOSAS", Product_line) ~ "Soft Drink",
  grepl("REFRESCOS", Product_line) ~ "Soft Drink",
  grepl("VINO", Product_line) ~ "Wine",
  grepl("ZUMO", Product_line) ~ "Soft Drink",
  Product_line == "ALIMENTACION" ~ "Food",
  Product_line == "BATIDOS" ~ "Soft Drink",
  Product_line == "BOTELLEROS" ~ "Specials",
  Product_line == "PLV" ~ "Specials",
  Product_line == "CAFE" ~ "Coffee",
  Product_line == "CO2" ~ "Soft Drink",
  Product_line == "LACTEOS" ~ "Food",
  Product_line == "LICORES" ~ "Liqueurs",
  Product_line == "LIMPIEZA" ~ "Specials",
  Product_line == "NAVIDAD" ~ "Specials",
  Product_line == "NO EXISTENCIAS" ~ "Specials",
))
```

As seen from above, the products are categorized into 9 types. This classification would improve our analysis to have a more in depth perspective to the data frame.

```
#select required variables
products <- ddi %>% select(Store_name,Product_line, Product_category)

#Only beer or soft drinks
products %>%
  group_by(Store_name) %>%
  filter(all(Product_category == "Beer" | Product_category == "Soft Drink")) %>%
  pull(Store_name) %>%
  n_distinct()
```

```
## [1] 298
```

We see that among 3199 stores in the dataset only **298** of them only buy beer or soft drinks.

```r
#not beer or soft drinks
products %>%
  group_by(Store_name) %>%
  filter(all(Product_category != "Beer" & Product_category != "Soft Drink")) %>%
  pull(Store_name) %>%
  n_distinct()
```

```
## [1] 204
```

Also among 3199 stores there are **204** stores which don't buy beer and soft drinks from DDI.

3.2 Challenge 2

This challenge will show how much wine or coffee can be sold in the designated areas. We won't use the prices that DDI gave us in the dataset, but we need to assume that a bar or restaurant can multiply the costs of these products. For wine this means 3 times a bottle and for coffee 10 times per kilo. We will show the results in the number of cases, bottles, kilos or euros.

Which data did we use? We again used the data of the categories of the products. Only this time for wine and coffee. To get the costs, and quantities we used the columns of the years and months. In the end we needed the area codes so that we could see the results for every designated area.

First we have to select the columns from the dataset that we need to use for this challenge. After we have done this we want to group them by their postal code and product category. This way we can summarize the results

```r
#getting wine & coffee
winecoffee <- ddi %>%
  select(Postal_Code, Product_category, Total_units20,
         Total_units19, Total_amount20, Total_amount19,
         Total_revenue20, Total_revenue19) %>%
 filter(Product_category == "Wine" | Product_category == "Coffee")
```

```r
#summarize based on each area and product and totals
winecoffee_group <- winecoffee %>%
  group_by(Postal_Code, Product_category) %>%
  summarise(Total_units = sum(Total_units20 + Total_units19),
            Total_amounts = sum(Total_amount20 + Total_amount19),
            Total_revenue = sum(Total_revenue20 + Total_revenue19))
```

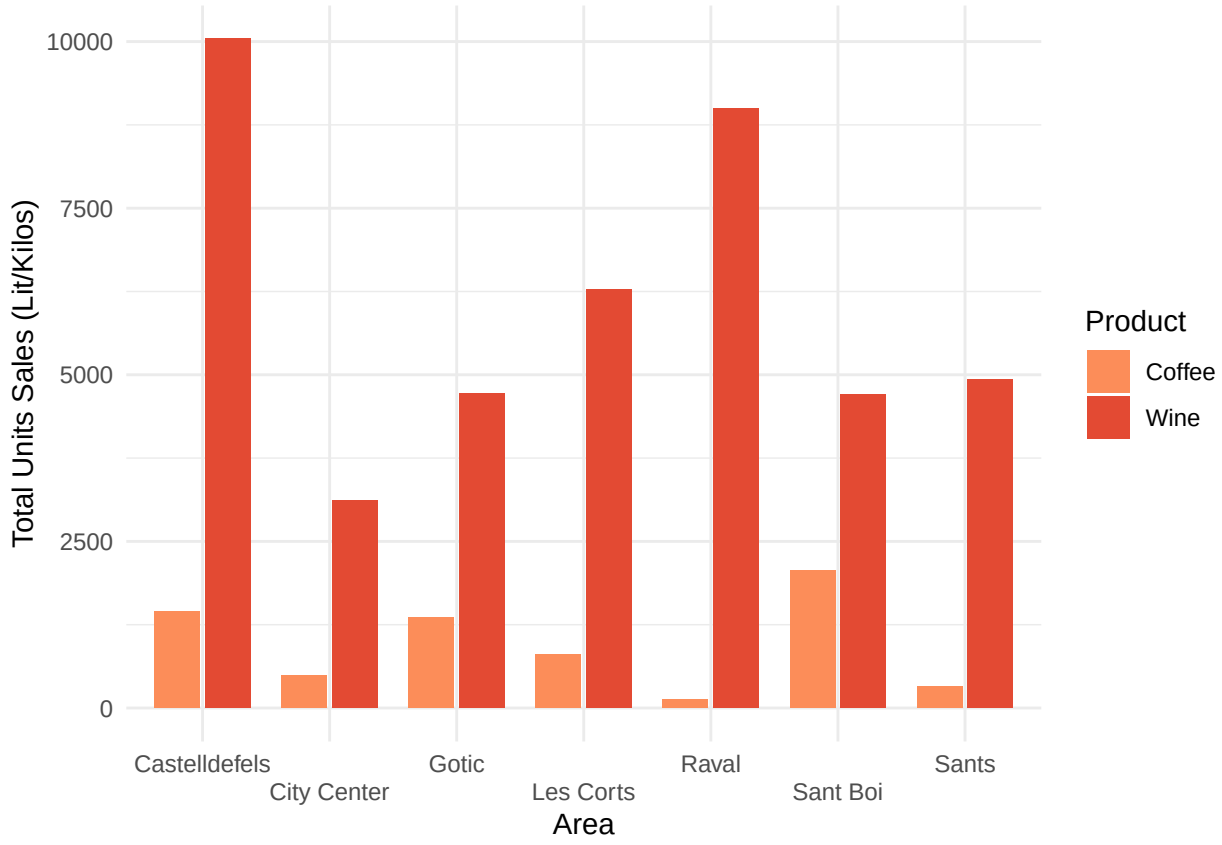
```
## `summarise()` has grouped output by 'Postal_Code'. You can override using the `.groups` argument.
```

```r
kable(winecoffee_group) %>%
kable_styling(bootstrap_options = "hover", full_width = F)
```

In this table we can see the total amount of units of wine or coffee sold to establishments. We can also see the amount of revenue that came in for DDI, selling these two products to the establishments. To get a better overview of the sales in kilos and liters, we made a bar graph of it so that we can compare the results per region.

Postal_Code	Product_category	Total_units	Total_amounts	Total_revenue
Castelldefels	Coffee	1448	1750	21404.99
Castelldefels	Wine	10039	51514	161476.16
City Center	Coffee	495	648	6575.10
City Center	Wine	3113	20990	44418.32
Gotic	Coffee	1361	1394	20038.76
Gotic	Wine	4722	28682	57600.80
Les Corts	Coffee	802	1085	13453.82
Les Corts	Wine	6287	40058	80301.49
Raval	Coffee	121	169	517.74
Raval	Wine	9000	68866	112245.22
Sant Boi	Coffee	2062	2125	29197.23
Sant Boi	Wine	4699	24272	75076.39
Sants	Coffee	325	593	4872.75
Sants	Wine	4926	28242	59229.69

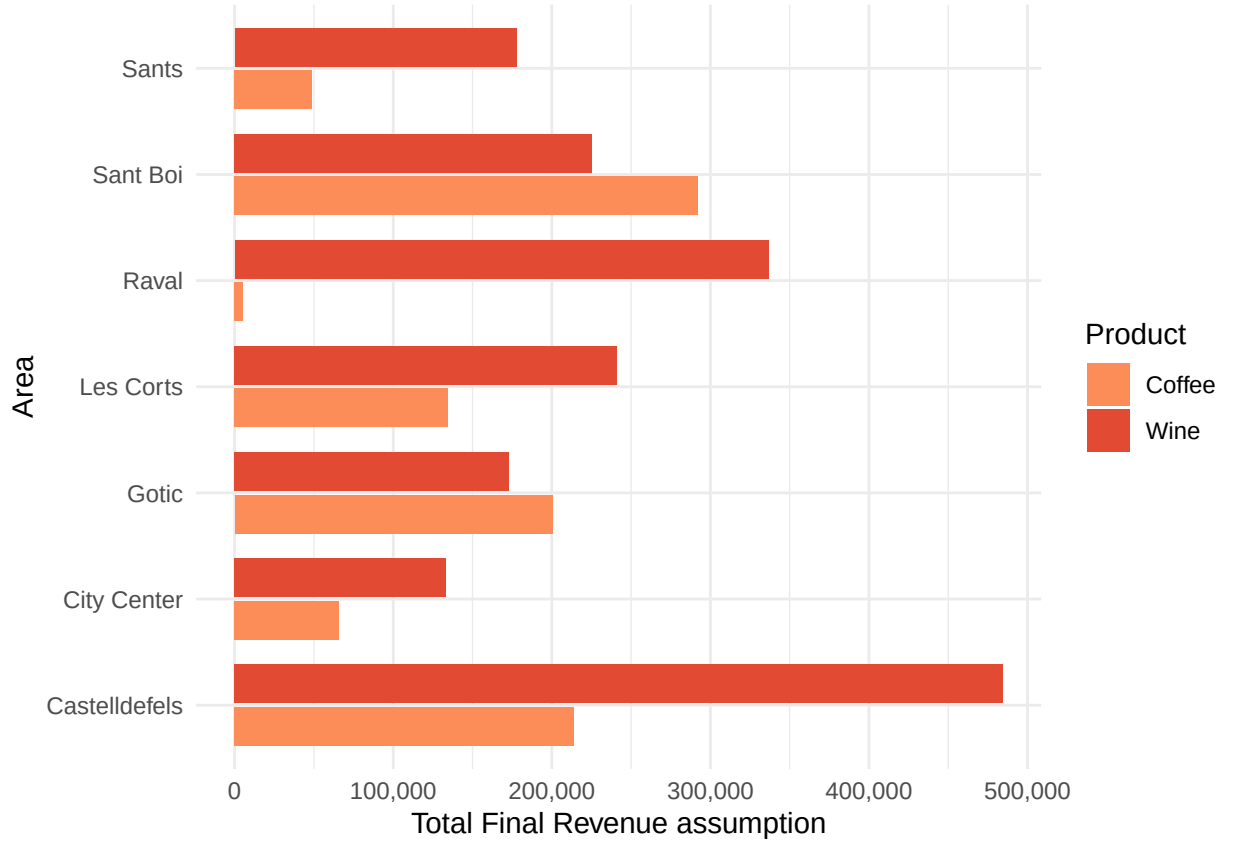
```
winecoffee_group %>%
  ggplot(aes(Postal_Code, Total_units, fill=Product_category)) +
  geom_col(position=position_dodge2(preserve = "single"), width=0.8) +
  scale_fill_manual(values = my_colors) +
  labs(x="Area", y= "Total Units Sales (Lit/Kilos)", fill= "Product") +
  guides(x = guide_axis(n.dodge = 2)) +
  theme_minimal()
```



```r
Final_revenue_pred <- winecoffee_group %>%
  mutate(Total_final_revenue =
         case_when(Product_category == "Wine" ~ (Total_revenue*3),
                   Product_category == "Coffee" ~ (Total_revenue*10)))

Final_revenue_pred
```

```
## # A tibble: 14 x 6
## # Groups:   Postal_Code [7]
##    Postal_Code   Product_category Total_units Total_amounts Total_revenue
##    <chr>         <chr>                 <dbl>         <dbl>         <dbl>
##  1 Castelldefels Coffee                 1448          1750        21405.
##  2 Castelldefels Wine                  10039         51514       161476.
##  3 City Center   Coffee                   495           648         6575.
##  4 City Center   Wine                    3113         20990        44418.
##  5 Gotic         Coffee                  1361          1394        20039.
##  6 Gotic         Wine                    4722         28682        57601.
##  7 Les Corts     Coffee                   802          1085        13454.
##  8 Les Corts     Wine                    6287         40058        80301.
##  9 Raval         Coffee                   121           169          518.
## 10 Raval         Wine                    9000         68866       112245.
## 11 Sant Boi      Coffee                  2062          2125        29197.
## 12 Sant Boi      Wine                    4699         24272        75076.
## 13 Sants         Coffee                   325           593         4873.
## 14 Sants         Wine                    4926         28242        59230.
## # ... with 1 more variable: Total_final_revenue <dbl>
```

```r
Final_revenue_pred %>%
  ggplot(aes(Postal_Code, Total_final_revenue, fill=Product_category)) +
  geom_col(position=position_dodge2(preserve = "single"), width=0.8) +
  scale_fill_manual(values = my_colors) +
  coord_flip() +
  scale_y_continuous(labels = label_comma()) +
  labs(x="Area", y= "Total Final Revenue assumption", fill= "Product") +
  theme_minimal()
```



3.3 Challenge 3

In this challenge we take a look at how many units of soft drinks, wine and coffee we sell in as compared to selling one unit of beer.

Which data did we use? From the data we again used the different product categories to compare the amount of units sold of beer, soft drinks, wine and coffee. To get the total amount of units sold for beer, coffee, wine and soft drinks we need to filter the product categories. After we have done this we need to group them by these categories so that we can count them together.

```
units_count <-
  ddi %>% select(Product_category, Total_units20, Total_units19) %>%
  filter(Product_category == c("Beer","Coffee","Wine","Soft Drink")) %>%
  group_by(Product_category) %>%
  summarise(Total_units20 = sum(Total_units20), Total_units19 = sum(Total_units19))


units_count
```

```
## # A tibble: 4 x 3
##   Product_category Total_units20 Total_units19
##   <chr>                    <dbl>         <dbl>
## 1 Beer                     91553        181740
## 2 Coffee                     754          1028
## 3 Soft Drink               59359        145036
## 4 Wine                      3511          7313
```

In 2019, we can see that 181740 units of beer were sold, 1028 units of coffee were sold, 145036 units of soft drinks were sold and 7313 units of wine were sold. So for each unit of beer sold, 0.00566 units of coffee, 0.80 units of soft drinks and 0.04 units of wine were sold respectively.

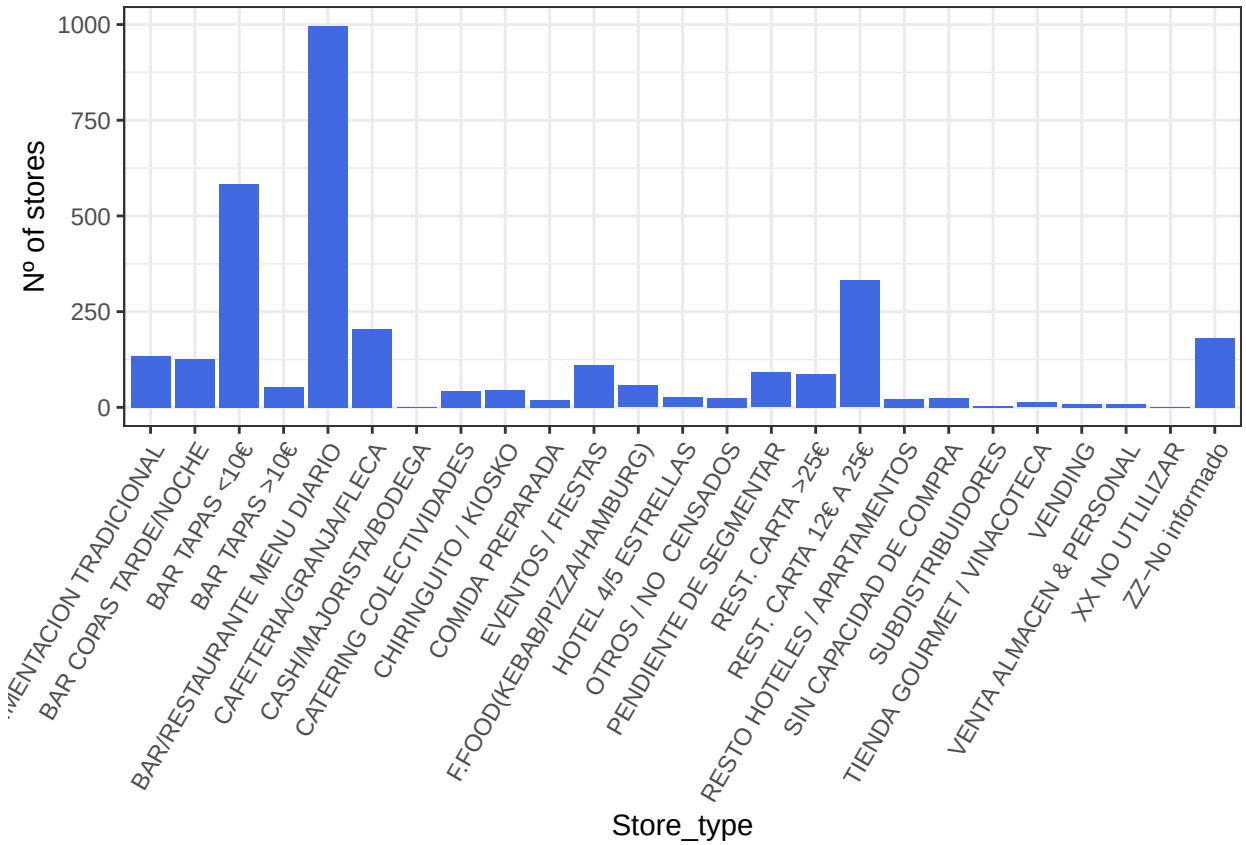
Although the amount of units in 2020 is not representative for a normal year because of covid-19 we still took a look at the number of units per each unit of beer sold.

4 Exploratory Analysis

Some graphs showing the stores

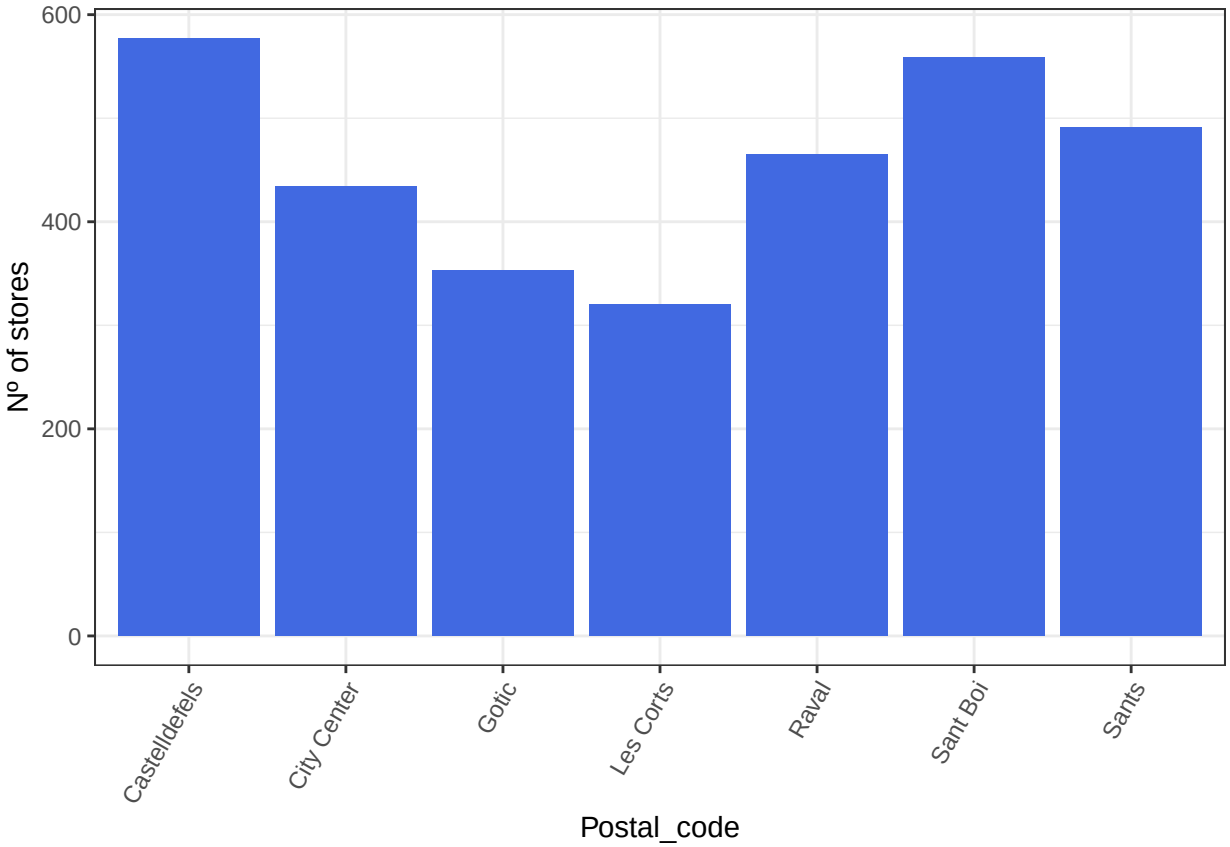
```
stores<- ddi %>%
  select(Postal_Code,Store_type,Store_name)


storestype <- stores %>%
    group_by(Store_type) %>%
    pull(Store_type)

stores_type<-stores%>%
    group_by(Store_type) %>%
  summarise(n_distinct(Store_name))
```

```
stores_type %>%
ggplot(aes(x=Store_type,y=`n_distinct(Store_name)`))   +
  geom_bar(stat = "Identity", fill="#4169E1") +
  labs(x= "Store_type", y = "Nº of stores")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



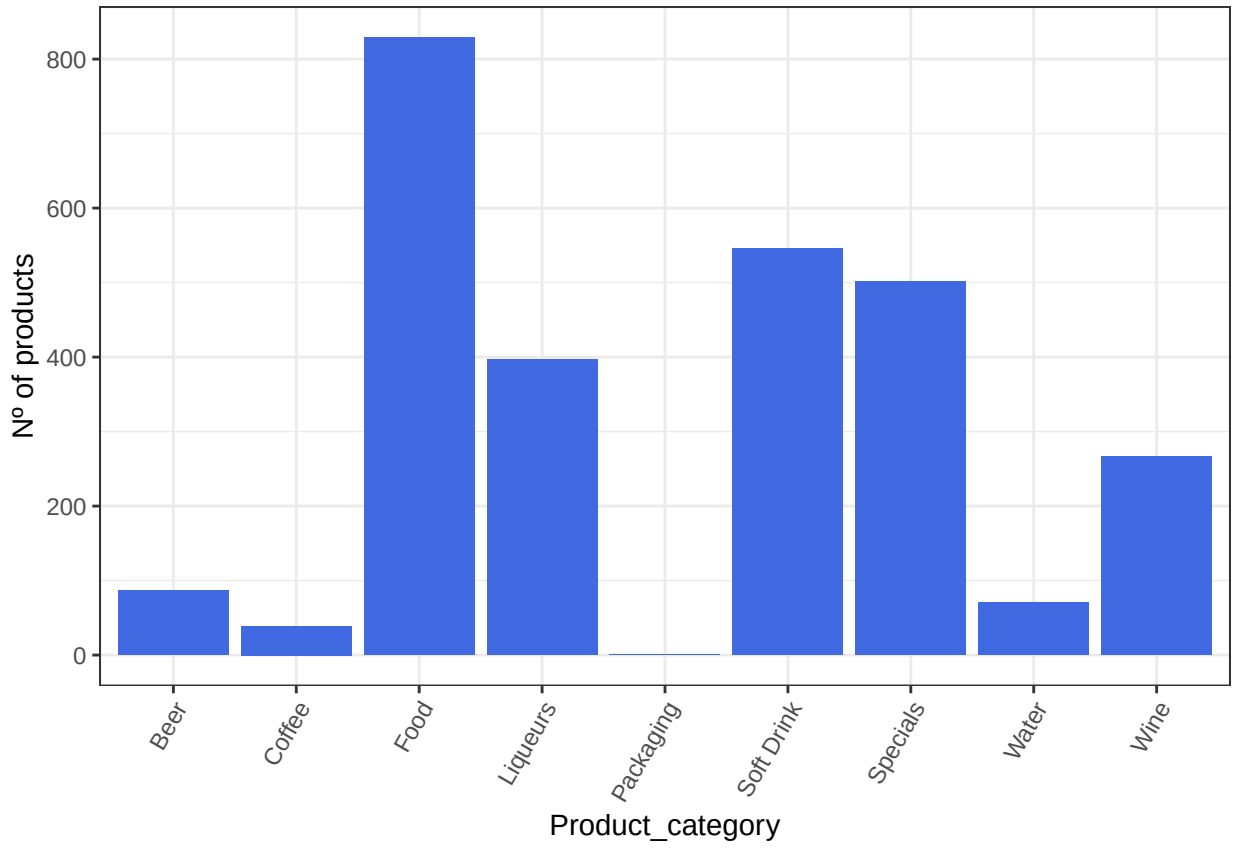

```
storesarea <- stores %>%
    group_by(Postal_Code) %>%
    pull(Postal_Code)

stores_area<-stores%>%
    group_by(Postal_Code) %>%
  summarise(n_distinct(Store_name))
```

```
stores_area %>%
ggplot(aes(x=Postal_Code,y=`n_distinct(Store_name)`))   +
  geom_bar(stat = "Identity", fill="#4169E1") +
  labs(x= "Postal_code", y = "Nº of stores")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

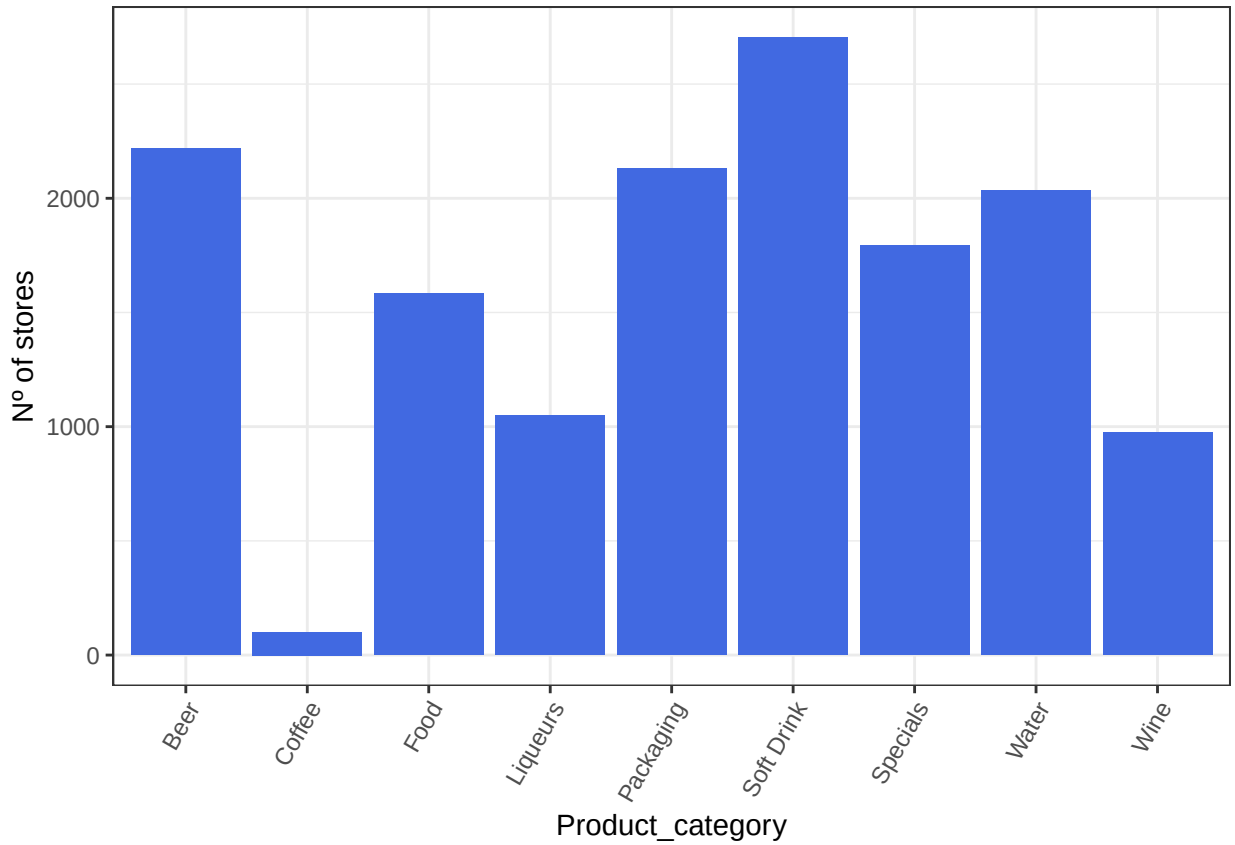


Some graphs showing the products

```
products<- ddi %>%
  select(Product_category,Product_name,Store_name)%>%
  group_by(Product_category) %>%
  summarise(n_distinct(Product_name),)

products %>%
ggplot(aes(x=Product_category,y=`n_distinct(Product_name)`))    +
  geom_bar(stat = "Identity", fill="#4169E1") +
  labs(x= "Product_category", y = "Nº of products")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```
products<- ddi %>%
  select(Product_category,Product_name,Store_name)%>%
  group_by(Product_category) %>%
  summarise(n_distinct(Store_name),)


products %>%
ggplot(aes(x=Product_category,y=`n_distinct(Store_name)`))   +
  geom_bar(stat = "Identity", fill="#4169E1") +
  labs(x= "Product_category", y = "Nº of stores")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

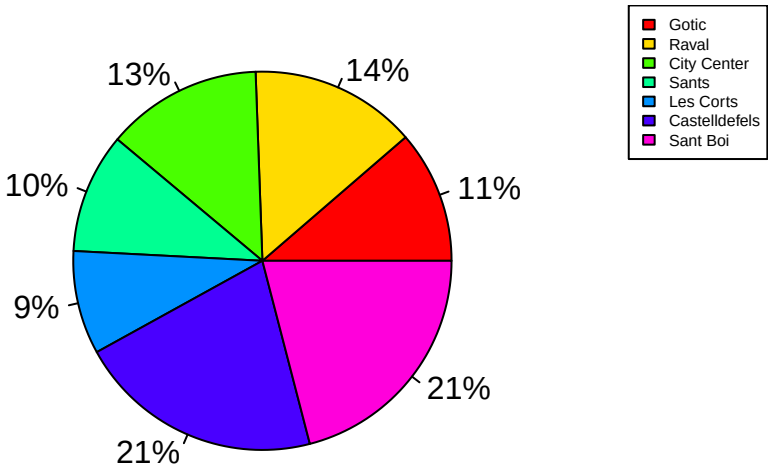


In the pie chart below you can see the difference in percentage between each region. It is good to see that there aren't any real lows or real highs. The lowest percentage is for the region Les Corts with 9% and the highest are for Castelldefels and Sant Boi with both 21%.

```
slices <- c(4613905,5796771,5426162,4179315,3602648,8554301,8532598)
lbls <- c("08001", "08002", "08003", "08014", "08028", "08830", "08860")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing revenues per region",col = rainbow(length(lbls)))
legend("topright", c("Gotic", "Raval", "City Center", "Sants", "Les Corts", "Castelldefels", "Sant Boi"
   fill = rainbow(length(lbls)))
```




```
#I
```

```
 ##Dividing product categories per Region
products <- ddi %>% select(Postal_Code, Product_category, Total_units20, Total_units19, Total_amount20,
 filter(Product_category == "Wine" | Product_category == "Coffee" | Product_category == "Beer"| Product_
```

```
#summarize based on each area and product and totals
products_group <- products %>% group_by(Postal_Code, Product_category) %>% summarise(Total_units = sum(
```

```
## `summarise()` has grouped output by 'Postal_Code'. You can override using the `.groups` argument.
```

```
products_group
```

```
## # A tibble: 56 x 5
## # Groups:   Postal_Code [7]
##    Postal_Code   Product_category Total_units Total_amounts Total_revenue
##    <chr>         <chr>                 <dbl>         <dbl>         <dbl>
##  1 Castelldefels Beer                 187681       2552576      5154681.
##  2 Castelldefels Coffee                 1448          1750        21405.
##  3 Castelldefels Food                  73541        498940       557950.
##  4 Castelldefels Packaging                 0             0       -78731.
##  5 Castelldefels Soft Drink           204072       1661652      2826706.
##  6 Castelldefels Specials                996         24953      -636245.
##  7 Castelldefels Water                110422       1437521       525356.
```

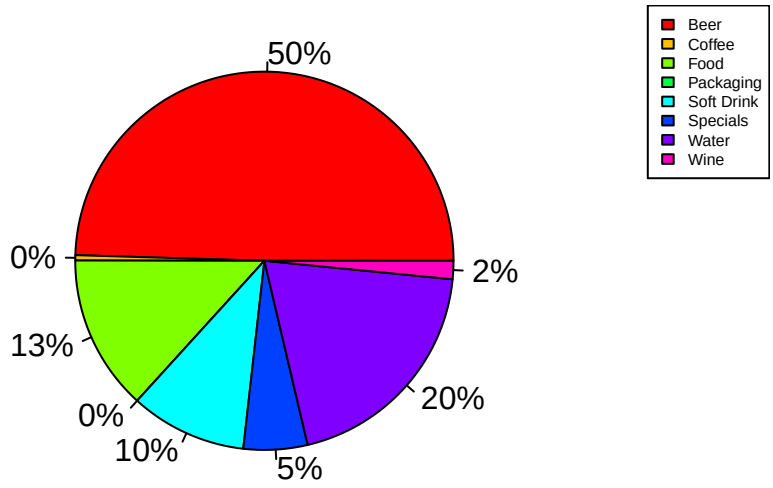
```
##  8 Castelldefels Wine                  10039         51514       161476.
##  9 City Center   Beer                  97418       1327885      2827826.
## 10 City Center   Coffee                  495           648         6575.
## # ... with 46 more rows
```

In the pie-chart below we can see that, in the region Gotic, by far the product that is sold the most is beer. Beer is 50% of the sales for the average establishment.

```
##Dividing product categories in Gotic
slices <- c(149737,1361,40155,0,30116,16510,59746,4722)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in Gotic",col = rainbow(length(lbls)))
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```



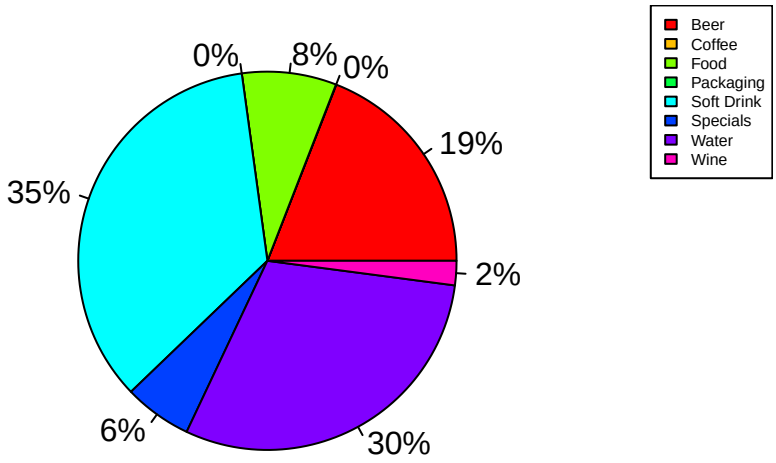

```
#
```

In the region of Raval we can see that it is a lot different than the pie-chart of Gotic. Instead of beer that was sold the most in Gotic, is it here the soft drinks that are sold the most with 35%. A close second here is water with 30%.

```
##Dividing product categories in Raval
slices <- c(82511,121,34940,0,151506,25121,129565,9000)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in Raval",col = rainbow(length(lbls)))
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```



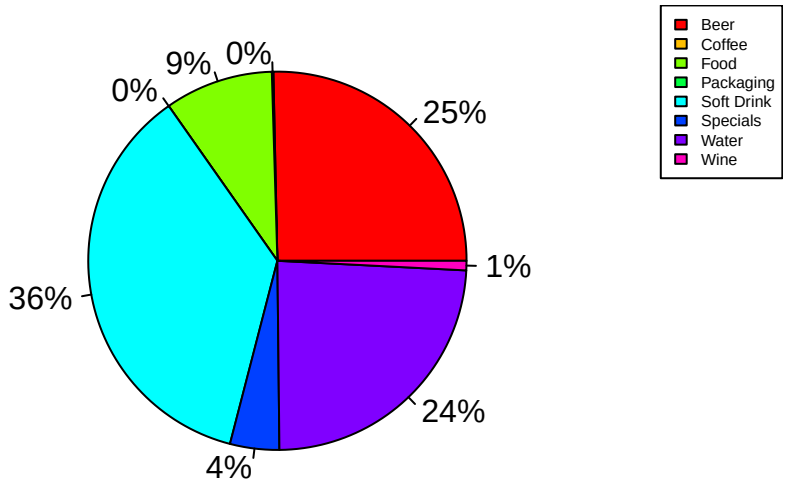

```
#
```

In the next pie-chart we see the results of the region City Center. Here is again one high with 36% of Soft Drinks. After this high there are another 2 big groups, beer and water with 25% and 24%.

```
##Dividing product categories in City Center
slices <- c(97418,495,35657,0,139237,16162,92336,3113)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in City Center",col = rainbow(length(lbls
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```



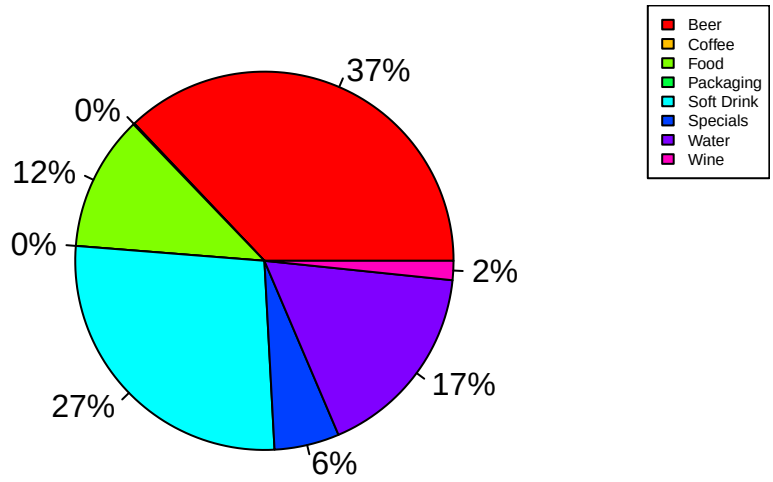

```
#
```

In the results of the region Sant, we can see that over of the sellings goes to beer. After this we can see that another big seller is soft drinks with 27%

```
##Dividing product categories in Sants
slices <- c(111361,325,34773,0,81442,16685,50933,4926)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in Sants",col = rainbow(length(lbls)))
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```



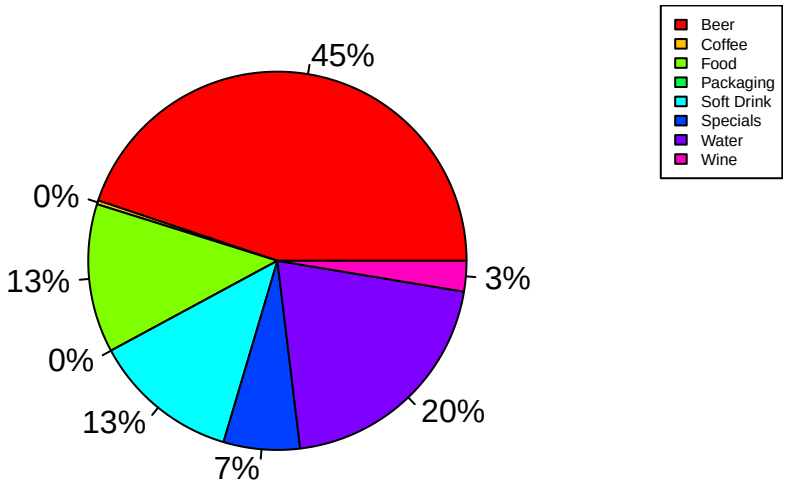

```
#
```

In Les Corts we can see again that beer is sold the most, with a respective 45%. Ather the beer, soft drinks are again the biggest with 20%.

```
##Dividing product categories in Les Corts
slices <- c(107865,802,30611,0,30114,15697,49233,6287)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in Les Corts",col = rainbow(length(lbls))
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```



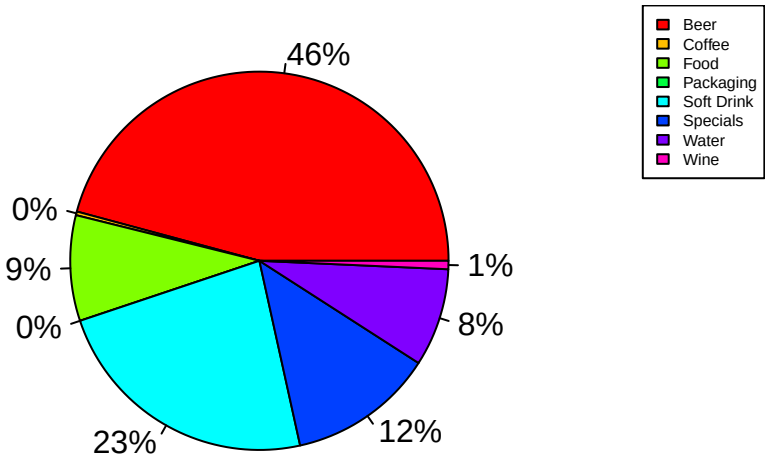

```
#
```

For the region Castelldefels, it is almost the same as for Les Corts. Again beer is the biggest product with 46% and the product after that is soft drinks with 23%.

```
##Dividing product categories in Castelldefels
slices <- c(301726,2062,59426,0,153595,82040,55166,4699)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in Castelldefels",col = rainbow(length(l
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```

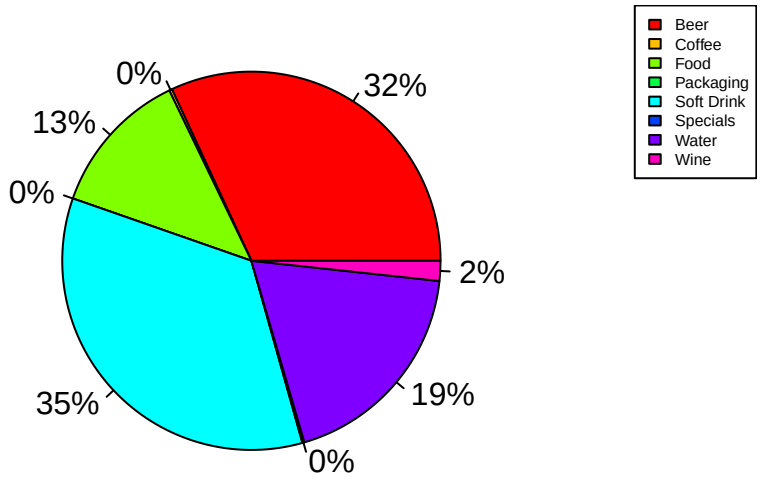


At last we have the region Sant Boi. This region has 2 major sellers, that are Soft drinks and beer with 35 and 32%.

```
##Dividing product categories in Sant Boi
slices <- c(187681,1448,73541,0,204072,996,110422,10039)
lbls <- c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine")
pct <- round(slices/sum(slices)*100)
lbls <- paste(pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels

slices +
  geom_text(aes(label = paste0(value,
                               " (",
                               scales::percent(value / sum(value)),
                               ")")),
            position = position_stack(vjust = 0.5))
```

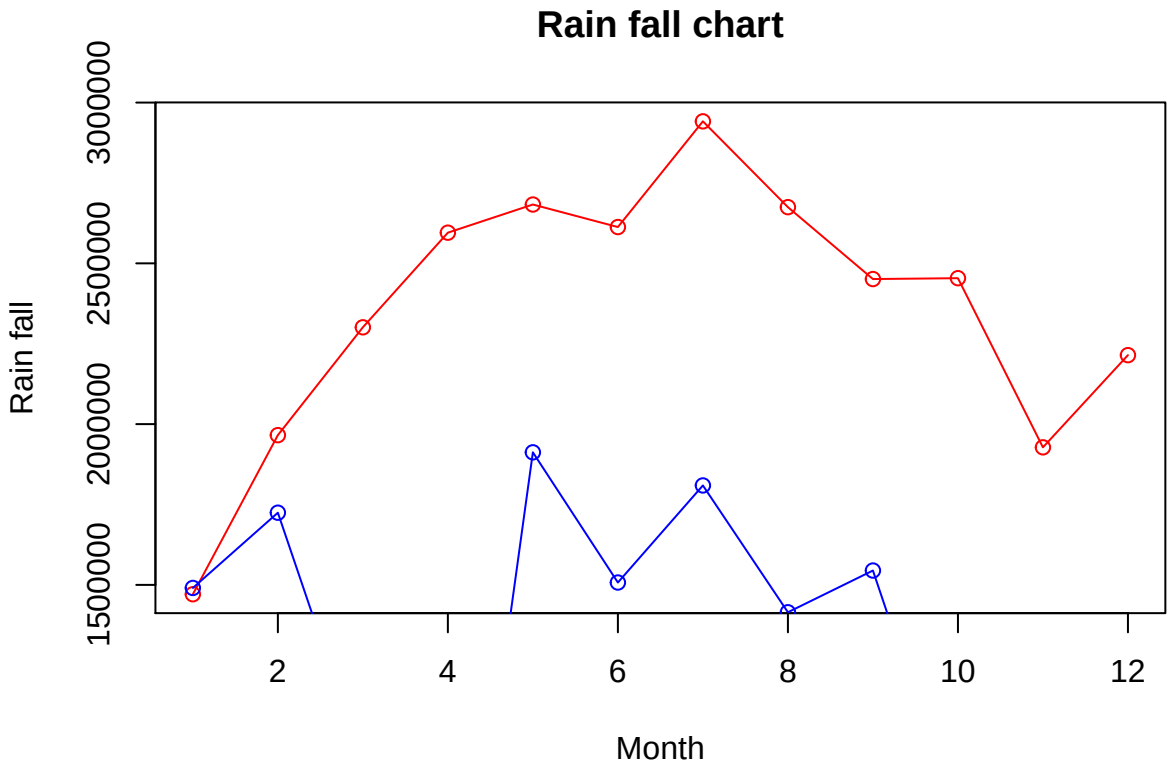
```
## NULL
```

```
pie(slices, labels = lbls, main = "Dividing product categories in Sant Boi",col = rainbow(length(lbls)))
legend("topright", c("Beer", "Coffee", "Food", "Packaging", "Soft Drink", "Specials", "Water", "Wine"),
   fill = rainbow(length(lbls)))
```



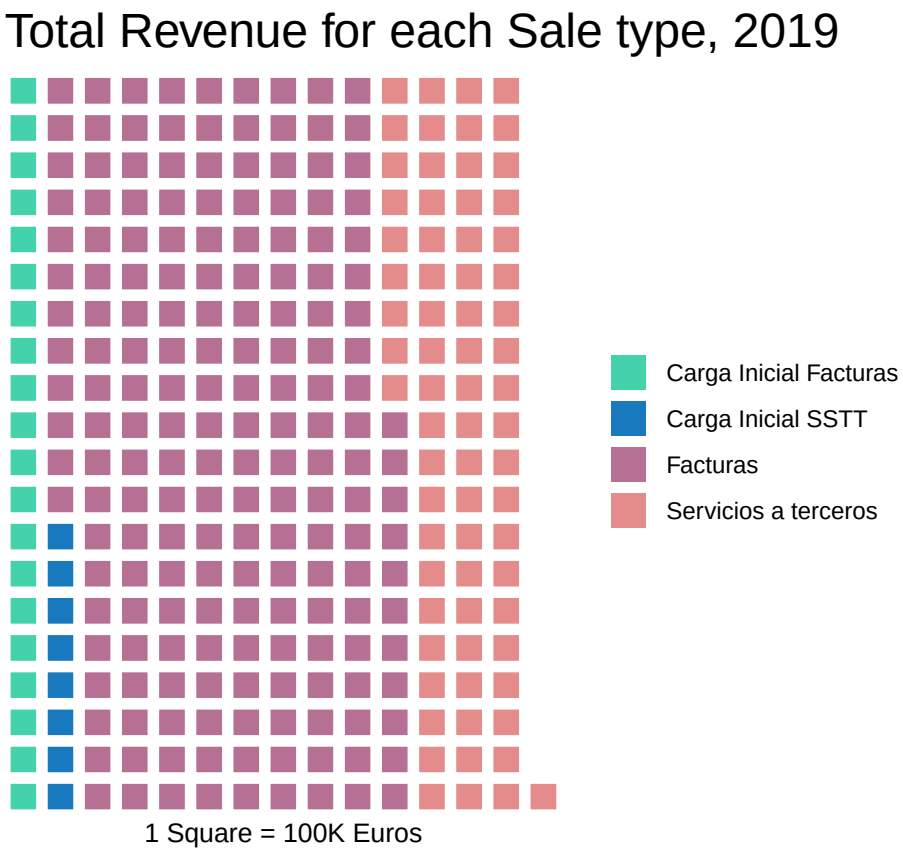
Linegraph with revenue streams each month in 2019 and 2020

```
v <- c(1470911,1965656,2301039,2595378,2682937,2612859,2941578,2674732,2451111,2453772,1927779,2214496)
t <- c(1490467,1724521,946281,-523,1912220,1507345,1809210,1414827,1544616,759651,428755,1126668)


# Plot the bar chart.
plot(v,type = "o",col = "red", xlab = "Month", ylab = "Rain fall",
   main = "Rain fall chart")

lines(t, type = "o", col = "blue")
```



Waffle graph for revenues based on each Sales type for

```
sales <- c(`Carga Inicial Facturas`=2041396.6, `Carga Inicial SSTT`=838946.1,
`Facturas`=18392255.6, `Servicios a terceros`=7019653.9)

waffle(sales/100000, rows=20,colors=c("#44D2AC", "#1879bf", "#B67093",
                "#E48B8B"), title="Total Revenue for each Sale type, 2019",
xlab="1 Square = 100K Euros")
```



5 Recommendations

5.1 Customer analytics

Choose some columns about customers.

```
ddi$Postal_Code<-as.factor(ddi$Postal_Code)
ddi$Store_type<-as.factor(ddi$Store_type)
ddi$Store_name<-as.factor(ddi$Store_name)
ddi$Sale_type<-as.factor(ddi$Sale_type)

ddi_customers<-ddi%>%
  select(Postal_Code,Store_type,Store_name,Sale_type,Total_revenue20,Total_revenue20)%>%
  group_by(Postal_Code)
```

```
set.seed(1212)
ddi_customers_split<-initial_split(ddi_customers,prop = 0.75)
```

```
ddi_customers_recipe<-training(ddi_customers_split)%>%
  recipe(Store_type~.)%>%
  step_corr(all_numeric())%>%
  step_center(all_numeric(),-all_numeric())%>%
  step_scale(all_numeric(),-all_numeric())%>%
  prep()
```

To build a model using the train set

```
ddi_customers_ranger<-rand_forest(mode = "classification")%>%
  set_engine("ranger")
ddi_customers_ranger_workflow<-workflow()%>%
  add_recipe(ddi_customers_recipe)%>%
  add_model(ddi_customers_ranger)%>%
  fit(training(ddi_customers_split))
ddi_customers_ranger_workflow
```

```
## == Workflow [trained] ==========================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 3 Recipe Steps
##
## * step_corr()
## * step_center()
## * step_scale()
##
## -- Model -----------------------------------------------------------------------
## Ranger result
##
## Call:
##  ranger::ranger(x = maybe_data_frame(x), y = y, num.threads = 1,      verbose = FALSE, seed = sample
##
## Type:                             Probability estimation
## Number of trees:                  500
## Sample size:                      70936
## Number of independent variables:  4
## Mtry:                             2
## Target node size:                 10
## Variable importance mode:         none
## Splitrule:                        gini
## OOB prediction error (Brier s.):  0.3653911
```

To predict the train set

```
ddi_customers_pred_train<-ddi_customers_ranger_workflow%>%
  predict(training(ddi_customers_split))%>%
  bind_cols(training(ddi_customers_split))


ddi_customers_pred_train%>%
  conf_mat(truth=Store_type,estimate=.pred_class)
```

```
##                              Truth
## Prediction                    ALIMENTACION TRADICIONAL BAR COPAS TARDE/NOCHE
##   ALIMENTACION TRADICIONAL                         2447                     7
##   BAR COPAS TARDE/NOCHE                               0                   909
##   BAR TAPAS <10\200                                  53                   166
##   BAR TAPAS >10\200                                   0                     0
```

```
##   BAR/RESTAURANTE MENU DIARIO                        448                  1290
##   CAFETERIA/GRANJA/FLECA                               0                     0
##   CASH/MAJORISTA/BODEGA                                0                     0
##   CATERING COLECTIVIDADES                              0                     1
##   CHIRINGUITO / KIOSKO                                 1                     0
##   COMIDA PREPARADA                                     0                     0
##   EVENTOS / FIESTAS                                    3                     7
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                          0                     0
##   HOTEL 4/5 ESTRELLAS                                  0                     0
##   OTROS / NO  CENSADOS                                 0                     0
##   PENDIENTE DE SEGMENTAR                               0                     0
##   REST. CARTA >25\200                                  0                     0
##   REST. CARTA 12\200 A 25\200                         12                    34
##   RESTO HOTELES / APARTAMENTOS                         0                     0
##   SIN CAPACIDAD DE COMPRA                              2                     0
##   SUBDISTRIBUIDORES                                    0                     0
##   TIENDA GOURMET / VINACOTECA                          0                     0
##   VENDING                                              0                     0
##   VENTA ALMACEN & PERSONAL                             0                     0
##   XX NO UTLILIZAR                                      0                     0
##   ZZ-No informado                                      0                     0
##                             Truth
## Prediction                   BAR TAPAS <10\200 BAR TAPAS >10\200
##   ALIMENTACION TRADICIONAL                5              9
##   BAR COPAS TARDE/NOCHE                   0              0
##   BAR TAPAS <10\200                    9879            116
##   BAR TAPAS >10\200                       0            505
##   BAR/RESTAURANTE MENU DIARIO          3548            585
##   CAFETERIA/GRANJA/FLECA                  8              1
##   CASH/MAJORISTA/BODEGA                   0              0
##   CATERING COLECTIVIDADES                 0              0
##   CHIRINGUITO / KIOSKO                    1              0
##   COMIDA PREPARADA                        0              0
##   EVENTOS / FIESTAS                      60              2
##   F.FOOD(KEBAB/PIZZA/HAMBURG)             0              0
##   HOTEL 4/5 ESTRELLAS                     0              0
##   OTROS / NO  CENSADOS                    0              0
##   PENDIENTE DE SEGMENTAR                  0              0
##   REST. CARTA >25\200                     0              1
##   REST. CARTA 12\200 A 25\200            52             16
##   RESTO HOTELES / APARTAMENTOS            2              0
##   SIN CAPACIDAD DE COMPRA                 0              0
##   SUBDISTRIBUIDORES                       0              0
##   TIENDA GOURMET / VINACOTECA             0              0
##   VENDING                                 0              0
##   VENTA ALMACEN & PERSONAL                0              0
##   XX NO UTLILIZAR                         0              0
##   ZZ-No informado                         0              0
##                             Truth
## Prediction                   BAR/RESTAURANTE MENU DIARIO
##   ALIMENTACION TRADICIONAL                            23
##   BAR COPAS TARDE/NOCHE                                2
##   BAR TAPAS <10\200                                  258
##   BAR TAPAS >10\200                                    0
```

```
##   BAR/RESTAURANTE MENU DIARIO                          25147
##   CAFETERIA/GRANJA/FLECA                                  33
##   CASH/MAJORISTA/BODEGA                                    0
##   CATERING COLECTIVIDADES                                  0
##   CHIRINGUITO / KIOSKO                                     0
##   COMIDA PREPARADA                                         0
##   EVENTOS / FIESTAS                                       11
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                              0
##   HOTEL 4/5 ESTRELLAS                                      0
##   OTROS / NO  CENSADOS                                     0
##   PENDIENTE DE SEGMENTAR                                   2
##   REST. CARTA >25\200                                      1
##   REST. CARTA 12\200 A 25\200                             48
##   RESTO HOTELES / APARTAMENTOS                             0
##   SIN CAPACIDAD DE COMPRA                                  0
##   SUBDISTRIBUIDORES                                        0
##   TIENDA GOURMET / VINACOTECA                              0
##   VENDING                                                  0
##   VENTA ALMACEN & PERSONAL                                 0
##   XX NO UTLILIZAR                                          0
##   ZZ-No informado                                        121
##                              Truth
## Prediction                    CAFETERIA/GRANJA/FLECA CASH/MAJORISTA/BODEGA
##   ALIMENTACION TRADICIONAL                        18                     0
##   BAR COPAS TARDE/NOCHE                            4                     0
##   BAR TAPAS <10\200                               98                     0
##   BAR TAPAS >10\200                                0                     0
##   BAR/RESTAURANTE MENU DIARIO                   2597                     1
##   CAFETERIA/GRANJA/FLECA                        1756                     0
##   CASH/MAJORISTA/BODEGA                            0                    27
##   CATERING COLECTIVIDADES                          0                     0
##   CHIRINGUITO / KIOSKO                             0                     0
##   COMIDA PREPARADA                                 0                     0
##   EVENTOS / FIESTAS                               12                     0
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                      0                     0
##   HOTEL 4/5 ESTRELLAS                              0                     0
##   OTROS / NO  CENSADOS                             0                     0
##   PENDIENTE DE SEGMENTAR                           2                     0
##   REST. CARTA >25\200                              0                     0
##   REST. CARTA 12\200 A 25\200                     11                     0
##   RESTO HOTELES / APARTAMENTOS                     0                     0
##   SIN CAPACIDAD DE COMPRA                          0                     0
##   SUBDISTRIBUIDORES                                0                     0
##   TIENDA GOURMET / VINACOTECA                      0                     0
##   VENDING                                          0                     0
##   VENTA ALMACEN & PERSONAL                         0                     0
##   XX NO UTLILIZAR                                  0                     0
##   ZZ-No informado                                 22                     0
##                              Truth
## Prediction                    CATERING COLECTIVIDADES CHIRINGUITO / KIOSKO
##   ALIMENTACION TRADICIONAL                         0                    17
##   BAR COPAS TARDE/NOCHE                            0                     0
##   BAR TAPAS <10\200                              133                    22
##   BAR TAPAS >10\200                                1                     0
```

```
##   BAR/RESTAURANTE MENU DIARIO                      181                 461
##   CAFETERIA/GRANJA/FLECA                             0                   0
##   CASH/MAJORISTA/BODEGA                              0                   0
##   CATERING COLECTIVIDADES                          436                   0
##   CHIRINGUITO / KIOSKO                               0                 433
##   COMIDA PREPARADA                                   0                   0
##   EVENTOS / FIESTAS                                  0                   0
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                        0                   0
##   HOTEL 4/5 ESTRELLAS                                2                   0
##   OTROS / NO  CENSADOS                               0                   0
##   PENDIENTE DE SEGMENTAR                             0                   0
##   REST. CARTA >25\200                                0                   0
##   REST. CARTA 12\200 A 25\200                       17                   2
##   RESTO HOTELES / APARTAMENTOS                       1                   0
##   SIN CAPACIDAD DE COMPRA                            0                   0
##   SUBDISTRIBUIDORES                                  0                   0
##   TIENDA GOURMET / VINACOTECA                        0                   0
##   VENDING                                            0                   0
##   VENTA ALMACEN & PERSONAL                           0                   0
##   XX NO UTLILIZAR                                    0                   0
##   ZZ-No informado                                    0                   2
##                              Truth
## Prediction                    COMIDA PREPARADA EVENTOS / FIESTAS
##   ALIMENTACION TRADICIONAL                   8                11
##   BAR COPAS TARDE/NOCHE                      0                 0
##   BAR TAPAS <10\200                         29               154
##   BAR TAPAS >10\200                          0                 0
##   BAR/RESTAURANTE MENU DIARIO               56               310
##   CAFETERIA/GRANJA/FLECA                    22                 1
##   CASH/MAJORISTA/BODEGA                      0                 0
##   CATERING COLECTIVIDADES                    0                 0
##   CHIRINGUITO / KIOSKO                       0                 0
##   COMIDA PREPARADA                          58                 0
##   EVENTOS / FIESTAS                          0              1810
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                0                 0
##   HOTEL 4/5 ESTRELLAS                        0                 0
##   OTROS / NO  CENSADOS                       0                 0
##   PENDIENTE DE SEGMENTAR                     1                 0
##   REST. CARTA >25\200                        0                 0
##   REST. CARTA 12\200 A 25\200               20                35
##   RESTO HOTELES / APARTAMENTOS               0                 0
##   SIN CAPACIDAD DE COMPRA                    0                 0
##   SUBDISTRIBUIDORES                          0                 0
##   TIENDA GOURMET / VINACOTECA                0                 0
##   VENDING                                    0                 0
##   VENTA ALMACEN & PERSONAL                   0                 0
##   XX NO UTLILIZAR                            0                 0
##   ZZ-No informado                            0                 0
##                              Truth
## Prediction                    F.FOOD(KEBAB/PIZZA/HAMBURG) HOTEL 4/5 ESTRELLAS
##   ALIMENTACION TRADICIONAL                              5                   0
##   BAR COPAS TARDE/NOCHE                                 0                   2
##   BAR TAPAS <10\200                                    40                   1
##   BAR TAPAS >10\200                                     3                   0
```

```
##   BAR/RESTAURANTE MENU DIARIO                          332                  71
##   CAFETERIA/GRANJA/FLECA                                 0                   0
##   CASH/MAJORISTA/BODEGA                                  0                   0
##   CATERING COLECTIVIDADES                                0                   0
##   CHIRINGUITO / KIOSKO                                   0                   0
##   COMIDA PREPARADA                                       0                   0
##   EVENTOS / FIESTAS                                     10                   9
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                          195                   0
##   HOTEL 4/5 ESTRELLAS                                    0                 107
##   OTROS / NO  CENSADOS                                   0                   0
##   PENDIENTE DE SEGMENTAR                                 0                   0
##   REST. CARTA >25\200                                    0                   0
##   REST. CARTA 12\200 A 25\200                           41                  12
##   RESTO HOTELES / APARTAMENTOS                           0                   3
##   SIN CAPACIDAD DE COMPRA                                0                   0
##   SUBDISTRIBUIDORES                                      0                   0
##   TIENDA GOURMET / VINACOTECA                            0                   0
##   VENDING                                                0                   0
##   VENTA ALMACEN & PERSONAL                               0                   0
##   XX NO UTLILIZAR                                        0                   0
##   ZZ-No informado                                        8                   0
##                              Truth
## Prediction                    OTROS / NO  CENSADOS PENDIENTE DE SEGMENTAR
##   ALIMENTACION TRADICIONAL                          4                  16
##   BAR COPAS TARDE/NOCHE                             0                   1
##   BAR TAPAS <10\200                                20                  24
##   BAR TAPAS >10\200                                 0                   0
##   BAR/RESTAURANTE MENU DIARIO                     207                 220
##   CAFETERIA/GRANJA/FLECA                            0                   1
##   CASH/MAJORISTA/BODEGA                             0                   0
##   CATERING COLECTIVIDADES                           0                   0
##   CHIRINGUITO / KIOSKO                              0                   0
##   COMIDA PREPARADA                                  0                   0
##   EVENTOS / FIESTAS                                 0                   4
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                       0                   0
##   HOTEL 4/5 ESTRELLAS                               0                   7
##   OTROS / NO  CENSADOS                             45                   0
##   PENDIENTE DE SEGMENTAR                            0                 385
##   REST. CARTA >25\200                               0                   0
##   REST. CARTA 12\200 A 25\200                       3                  38
##   RESTO HOTELES / APARTAMENTOS                      0                   2
##   SIN CAPACIDAD DE COMPRA                           0                   0
##   SUBDISTRIBUIDORES                                 0                   0
##   TIENDA GOURMET / VINACOTECA                       0                   0
##   VENDING                                           0                   0
##   VENTA ALMACEN & PERSONAL                          0                   0
##   XX NO UTLILIZAR                                   0                   0
##   ZZ-No informado                                   0                  30
##                              Truth
## Prediction                    REST. CARTA >25\200 REST. CARTA 12\200 A 25\200
##   ALIMENTACION TRADICIONAL                    3                    20
##   BAR COPAS TARDE/NOCHE                       2                     1
##   BAR TAPAS <10\200                         181                   282
##   BAR TAPAS >10\200                           0                     0
```

```
##   BAR/RESTAURANTE MENU DIARIO              1068                  3151
##   CAFETERIA/GRANJA/FLECA                      0                    16
##   CASH/MAJORISTA/BODEGA                       0                     0
##   CATERING COLECTIVIDADES                     0                     0
##   CHIRINGUITO / KIOSKO                        0                     0
##   COMIDA PREPARADA                            0                     0
##   EVENTOS / FIESTAS                          29                    49
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                 1                     0
##   HOTEL 4/5 ESTRELLAS                         0                     0
##   OTROS / NO  CENSADOS                        0                     0
##   PENDIENTE DE SEGMENTAR                      0                     0
##   REST. CARTA >25\200                       812                     5
##   REST. CARTA 12\200 A 25\200                22                  4735
##   RESTO HOTELES / APARTAMENTOS                6                     0
##   SIN CAPACIDAD DE COMPRA                     0                     0
##   SUBDISTRIBUIDORES                           0                     0
##   TIENDA GOURMET / VINACOTECA                 0                     0
##   VENDING                                     0                     0
##   VENTA ALMACEN & PERSONAL                    0                     0
##   XX NO UTLILIZAR                             0                     0
##   ZZ-No informado                             0                    42
##                              Truth
## Prediction                    RESTO HOTELES / APARTAMENTOS
##   ALIMENTACION TRADICIONAL                                0
##   BAR COPAS TARDE/NOCHE                                   0
##   BAR TAPAS <10\200                                      29
##   BAR TAPAS >10\200                                       0
##   BAR/RESTAURANTE MENU DIARIO                            29
##   CAFETERIA/GRANJA/FLECA                                  0
##   CASH/MAJORISTA/BODEGA                                   0
##   CATERING COLECTIVIDADES                                 0
##   CHIRINGUITO / KIOSKO                                    0
##   COMIDA PREPARADA                                        0
##   EVENTOS / FIESTAS                                       4
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                             0
##   HOTEL 4/5 ESTRELLAS                                     1
##   OTROS / NO  CENSADOS                                    0
##   PENDIENTE DE SEGMENTAR                                  0
##   REST. CARTA >25\200                                     0
##   REST. CARTA 12\200 A 25\200                             9
##   RESTO HOTELES / APARTAMENTOS                          108
##   SIN CAPACIDAD DE COMPRA                                 0
##   SUBDISTRIBUIDORES                                       0
##   TIENDA GOURMET / VINACOTECA                             0
##   VENDING                                                 0
##   VENTA ALMACEN & PERSONAL                                0
##   XX NO UTLILIZAR                                         0
##   ZZ-No informado                                         5
##                              Truth
## Prediction                    SIN CAPACIDAD DE COMPRA SUBDISTRIBUIDORES
##   ALIMENTACION TRADICIONAL                          0                 0
##   BAR COPAS TARDE/NOCHE                             0                 0
##   BAR TAPAS <10\200                                16                 0
##   BAR TAPAS >10\200                                 0                 0
```

```
##   BAR/RESTAURANTE MENU DIARIO                        85              37
##   CAFETERIA/GRANJA/FLECA                              0               0
##   CASH/MAJORISTA/BODEGA                               0               0
##   CATERING COLECTIVIDADES                             0               0
##   CHIRINGUITO / KIOSKO                                0               0
##   COMIDA PREPARADA                                    0               0
##   EVENTOS / FIESTAS                                   1               0
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                         0               0
##   HOTEL 4/5 ESTRELLAS                                 0               0
##   OTROS / NO  CENSADOS                                0               0
##   PENDIENTE DE SEGMENTAR                              1               0
##   REST. CARTA >25\200                                 0               0
##   REST. CARTA 12\200 A 25\200                         0              10
##   RESTO HOTELES / APARTAMENTOS                        0               0
##   SIN CAPACIDAD DE COMPRA                            60               0
##   SUBDISTRIBUIDORES                                   0              38
##   TIENDA GOURMET / VINACOTECA                         0               0
##   VENDING                                             0               0
##   VENTA ALMACEN & PERSONAL                            0               0
##   XX NO UTLILIZAR                                     0               0
##   ZZ-No informado                                     3               0
##                             Truth
## Prediction                   TIENDA GOURMET / VINACOTECA VENDING
##   ALIMENTACION TRADICIONAL                             1       0
##   BAR COPAS TARDE/NOCHE                                0       1
##   BAR TAPAS <10\200                                   27       6
##   BAR TAPAS >10\200                                    0       0
##   BAR/RESTAURANTE MENU DIARIO                        163      56
##   CAFETERIA/GRANJA/FLECA                               0       0
##   CASH/MAJORISTA/BODEGA                                0       0
##   CATERING COLECTIVIDADES                              0       0
##   CHIRINGUITO / KIOSKO                                 0       0
##   COMIDA PREPARADA                                     0       0
##   EVENTOS / FIESTAS                                    0       3
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                          0       0
##   HOTEL 4/5 ESTRELLAS                                  0       0
##   OTROS / NO  CENSADOS                                 0       0
##   PENDIENTE DE SEGMENTAR                               0       0
##   REST. CARTA >25\200                                  0       0
##   REST. CARTA 12\200 A 25\200                          0       0
##   RESTO HOTELES / APARTAMENTOS                         0       0
##   SIN CAPACIDAD DE COMPRA                              0       0
##   SUBDISTRIBUIDORES                                    0       0
##   TIENDA GOURMET / VINACOTECA                          9       0
##   VENDING                                              0       0
##   VENTA ALMACEN & PERSONAL                             0       0
##   XX NO UTLILIZAR                                      0       0
##   ZZ-No informado                                      0       0
##                             Truth
## Prediction                   VENTA ALMACEN & PERSONAL XX NO UTLILIZAR
##   ALIMENTACION TRADICIONAL                           0              0
##   BAR COPAS TARDE/NOCHE                              0              0
##   BAR TAPAS <10\200                                  0              2
##   BAR TAPAS >10\200                                  0              0
```

```
##   BAR/RESTAURANTE MENU DIARIO                        40              10
##   CAFETERIA/GRANJA/FLECA                              0               0
##   CASH/MAJORISTA/BODEGA                               0               0
##   CATERING COLECTIVIDADES                             0               0
##   CHIRINGUITO / KIOSKO                                0               0
##   COMIDA PREPARADA                                    0               0
##   EVENTOS / FIESTAS                                   0               0
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                         0               0
##   HOTEL 4/5 ESTRELLAS                                 0               0
##   OTROS / NO  CENSADOS                                0               0
##   PENDIENTE DE SEGMENTAR                              1               0
##   REST. CARTA >25\200                                 0               0
##   REST. CARTA 12\200 A 25\200                         2               0
##   RESTO HOTELES / APARTAMENTOS                        0               0
##   SIN CAPACIDAD DE COMPRA                             0               0
##   SUBDISTRIBUIDORES                                   0               0
##   TIENDA GOURMET / VINACOTECA                         0               0
##   VENDING                                             0               0
##   VENTA ALMACEN & PERSONAL                           68               0
##   XX NO UTLILIZAR                                     0               0
##   ZZ-No informado                                     0               0
##                             Truth
## Prediction                   ZZ-No informado
##   ALIMENTACION TRADICIONAL                    0
##   BAR COPAS TARDE/NOCHE                       0
##   BAR TAPAS <10\200                           3
##   BAR TAPAS >10\200                           0
##   BAR/RESTAURANTE MENU DIARIO                71
##   CAFETERIA/GRANJA/FLECA                      0
##   CASH/MAJORISTA/BODEGA                       0
##   CATERING COLECTIVIDADES                     0
##   CHIRINGUITO / KIOSKO                        0
##   COMIDA PREPARADA                            0
##   EVENTOS / FIESTAS                           0
##   F.FOOD(KEBAB/PIZZA/HAMBURG)                 0
##   HOTEL 4/5 ESTRELLAS                         0
##   OTROS / NO  CENSADOS                        1
##   PENDIENTE DE SEGMENTAR                      0
##   REST. CARTA >25\200                         0
##   REST. CARTA 12\200 A 25\200                 1
##   RESTO HOTELES / APARTAMENTOS                0
##   SIN CAPACIDAD DE COMPRA                     0
##   SUBDISTRIBUIDORES                           0
##   TIENDA GOURMET / VINACOTECA                 0
##   VENDING                                     0
##   VENTA ALMACEN & PERSONAL                    0
##   XX NO UTLILIZAR                             0
##   ZZ-No informado                          3177
```

model performance

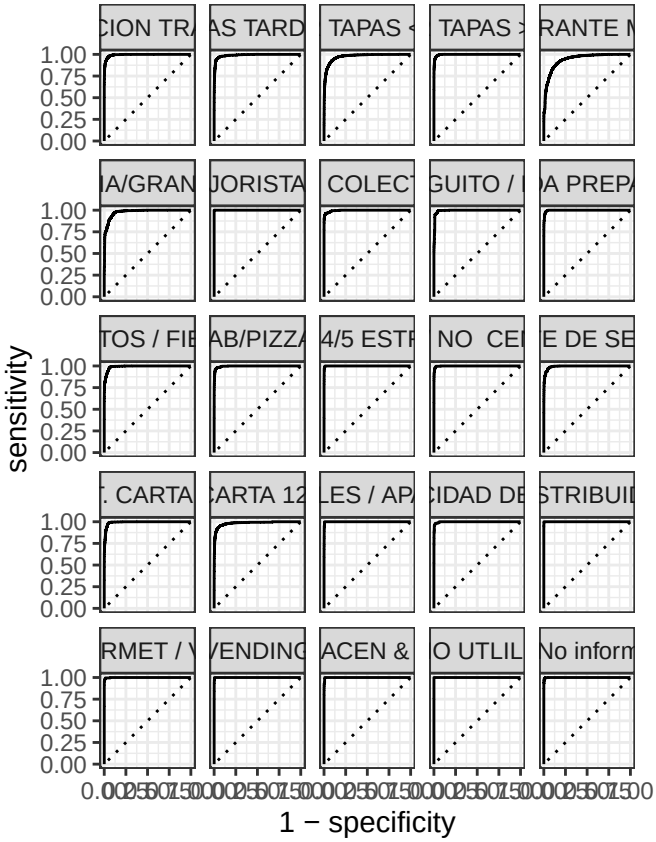
We can see that the accuracy of the model is 0.803,the presition of model is 0.961 and the real of model is 0.581.

ROC curve

```
ddi_customers_prod_train%>%
  roc_auc(Store_type,`.pred_ALIMENTACION TRADICIONAL`,`.pred_BAR COPAS TARDE/NOCHE`,`.pred_BAR TAPAS <1
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc hand_till      0.992
```

```
ddi_customers_prod_train%>%
  roc_curve(Store_type,`.pred_ALIMENTACION TRADICIONAL`,`.pred_BAR COPAS TARDE/NOCHE`,`.pred_BAR TAPAS <
  autoplot()
```



Test set:confusion matrics

```
ddi_customers_ranger_workflow%>%
  predict(testing(ddi_customers_split))%>%
  bind_cols(testing(ddi_customers_split))%>%
  class_metrics(truth=Store_type,estimate=.pred_class)
```

```
## # A tibble: 3 x 3
##   .metric   .estimator .estimate
##   <chr>     <chr>          <dbl>
## 1 accuracy  multiclass     0.705
## 2 precision macro          0.907
## 3 recall    macro          0.444
```

5.2 Basket Analysis

We want to understand which products are bought more regarding frequency in order to have a product recommendation system for the distribution company. Since there are 2736 unique products, we will only concentrate on rules with more significance.

Building the transaction matrix Let's build the transaction table:

-as rows as transactions, -as columns as items, -each elements a logical variable indicating if item j is in transaction i.

```
wine <- ddi %>% filter(Product_category == "Wine") %>% select(Store_name, Product_name)
ddi_table <-
  wine %>%
  group_by(Store_name, Product_name) %>%
  summarise(exists = TRUE, .groups = "drop") %>%
  pivot_wider(names_from = "Product_name", values_from = "exists") %>%
  replace(is.na(.), FALSE)
```

The result is a table with 3199 Stores & 2736 Products.

From the transaction table we get the transaction matrix and after that apply the rules.

```
ddi_matrix <- as(ddi_table %>% select(-Store_name), "transactions")
```

5.2.1 Rules mining

When applying the condition for rules, we can target different result. In this case, we will target a lot of transactions by having a 3 percent support but on the other hand with a confidence of 85 percent to obtain the most frequent transactions.

```
rules_wine <- apriori(ddi_matrix,
                 control = list(verbose = FALSE),
                 parameter = list(supp = 0.01, conf = 0.30))

summary(rules_wine)
```

```
## set of 397 rules
##
## rule length distribution (lhs + rhs):sizes
##   2   3   4   5
##  77 238  73   9
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   3.000   3.035   3.000   5.000
##
## summary of quality measures:
##     support          confidence        coverage            lift
##  Min.   :0.01027   Min.   :0.3000   Min.   :0.01027   Min.   : 1.574
##  1st Qu.:0.01129   1st Qu.:0.4286   1st Qu.:0.01745   1st Qu.: 4.030
##  Median :0.01335   Median :0.6316   Median :0.02464   Median : 5.939
##  Mean   :0.01826   Mean   :0.6363   Mean   :0.03152   Mean   : 7.380
##  3rd Qu.:0.01745   3rd Qu.:0.8333   3rd Qu.:0.03388   3rd Qu.: 8.855
```

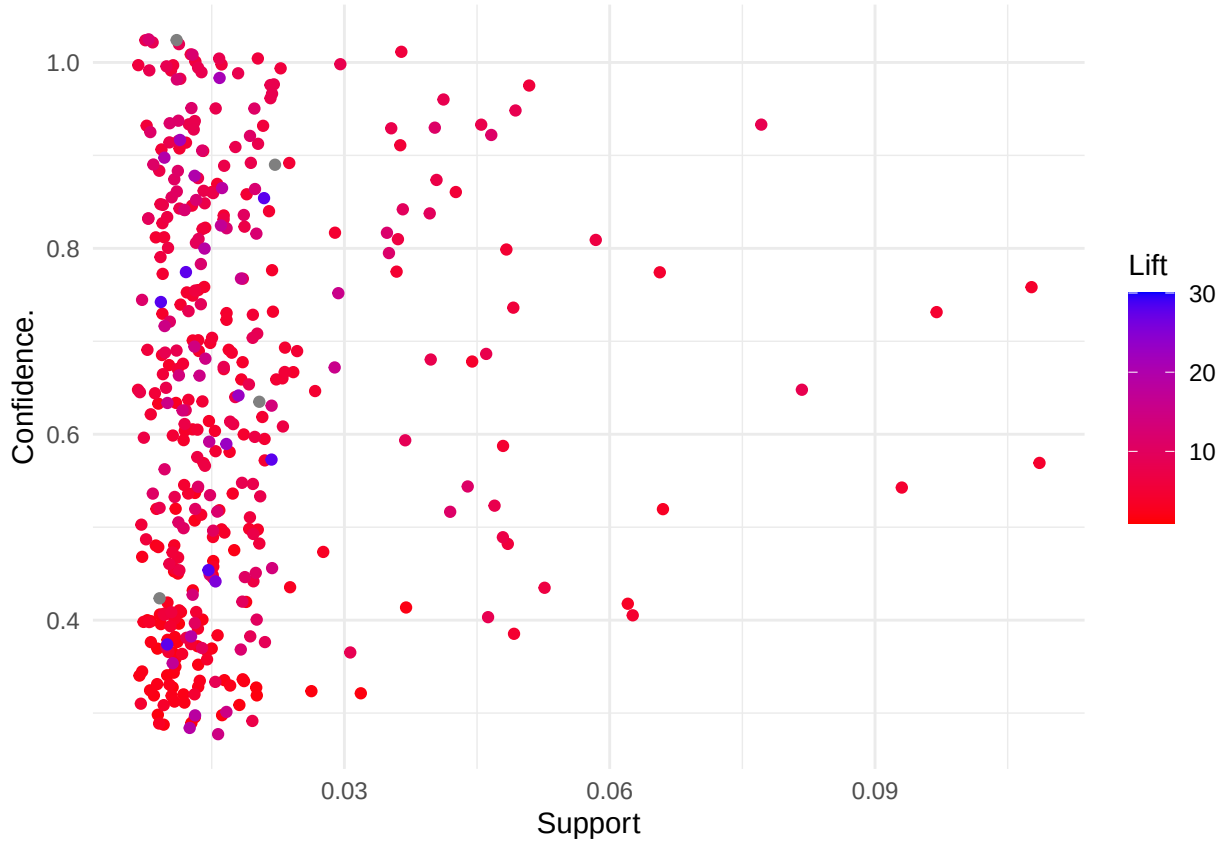
```
##  Max.   :0.10575   Max.   :1.0000   Max.   :0.18172   Max.   :32.467
##       count
##  Min.   : 10.00
##  1st Qu.: 11.00
##  Median : 13.00
##  Mean   : 17.78
##  3rd Qu.: 17.00
##  Max.   :103.00
##
## mining info:
##        data ntransactions support confidence
##  ddi_matrix           974    0.01        0.3
```

Greater lift values indicate stronger associations

```
#sort by lift
rules_wine <- sort(rules_wine, by="lift", decreasing = TRUE)
```

Scatter plot for all rules

```
plot(rules_wine, engine = "ggplot2", main = NULL, jitter = 2) +
scale_color_gradient2(low = "green", mid = "red", high = "blue",
midpoint = 1, limits = c(1,30)) +
labs(x = "Support", y = "Confidence.", color = "Lift") +
theme_minimal()
```



We see more density with lower Confidence and Support. As support increses the lift will get lower.

```
## display some rules
inspect(rules_wine[1:10], ruleSep = "~~>", itemSep = " + ", setStart = "", setEnd = "",
  linebreak = FALSE)
```

```
##      lhs
## [1]  CARLOS SERRES MAGNUM CR. - 501575        ~~>
## [2]  CARLOS SERRES TINTO 75CL 6U - 70999      ~~>
## [3]  TORREBLANCA ROSAT 6U - 72040             ~~>
## [4]  SANTANA BLANCO  75CL 12U - 71932         ~~>
## [5]  BERBERANA TAPAS ROSADO 6U - 504498       ~~>
## [6]  BERBERANA TAPAS BLANCO 6U - 70896        ~~>
## [7]  CASTELL SABINYA 75CL TINTO 12U - 503213  ~~>
## [8]  CASTELL SABINYA 75CL BLANCO 12U - 506569 ~~>
## [9]  UVA PIRATA PETIT VERDOT 6U - 400253      ~~>
## [10] OSTRAS PEDRIN VERDOSILLA 6U - 300273     ~~>
##      rhs                                      support    confidence coverage
## [1]  CARLOS SERRES TINTO 75CL 6U - 70999      0.01232033 1.0000000  0.01232033
## [2]  CARLOS SERRES MAGNUM CR. - 501575        0.01232033 0.4000000  0.03080082
## [3]  SANTANA BLANCO  75CL 12U - 71932         0.01848049 0.6428571  0.02874743
## [4]  TORREBLANCA ROSAT 6U - 72040             0.01848049 0.9000000  0.02053388
## [5]  BERBERANA TAPAS BLANCO 6U - 70896        0.01232033 0.7500000  0.01642710
## [6]  BERBERANA TAPAS ROSADO 6U - 504498       0.01232033 0.4615385  0.02669405
## [7]  CASTELL SABINYA 75CL BLANCO 12U - 506569 0.01848049 0.6000000  0.03080082
## [8]  CASTELL SABINYA 75CL TINTO 12U - 503213  0.01848049 0.8571429  0.02156057
## [9]  OSTRAS PEDRIN VERDOSILLA 6U - 300273     0.01026694 0.7692308  0.01334702
## [10] UVA PIRATA PETIT VERDOT 6U - 400253      0.01026694 0.3703704  0.02772074
##      lift     count
## [1]  32.46667 12
## [2]  32.46667 12
## [3]  31.30714 18
## [4]  31.30714 18
## [5]  28.09615 12
## [6]  28.09615 12
## [7]  27.82857 18
## [8]  27.82857 18
## [9]  27.74929 10
## [10] 27.74929 10
```

As the result show, we can see based on rule 8 that 96 percent of Stores who bought "LETONA CAMBIOS 1L" also bought "LETONA ENTERA GRAN CREM 1L RET 12U" product.

From this point the analysis can be very wide. For instance, sorting with lift would allow us to interpret the most significant rules. Below a specific product of **Soft Drink** category is analyzed to illustrate the method.

```
rules_ex <- subset(rules_wine, subset = lift > 25 )
inspect(rules_ex)
```

```
##      lhs                                      rhs                                      supp
## [1]  {CARLOS SERRES MAGNUM CR. - 501575}      => {CARLOS SERRES TINTO 75CL 6U - 70999} 0.0123
## [2]  {CARLOS SERRES TINTO 75CL 6U - 70999}    => {CARLOS SERRES MAGNUM CR. - 501575}   0.0123
## [3]  {TORREBLANCA ROSAT 6U - 72040}           => {SANTANA BLANCO  75CL 12U - 71932}    0.0184
```

```
## [4]  {SANTANA BLANCO  75CL 12U - 71932}         => {TORREBLANCA ROSAT 6U - 72040}             0.01848
## [5]  {BERBERANA TAPAS ROSADO 6U - 504498}       => {BERBERANA TAPAS BLANCO 6U - 70896}        0.01231
## [6]  {BERBERANA TAPAS BLANCO 6U - 70896}        => {BERBERANA TAPAS ROSADO 6U - 504498}       0.01231
## [7]  {CASTELL SABINYA 75CL TINTO 12U - 503213}  => {CASTELL SABINYA 75CL BLANCO 12U - 506569} 0.01848
## [8]  {CASTELL SABINYA 75CL BLANCO 12U - 506569} => {CASTELL SABINYA 75CL TINTO 12U - 503213}  0.01848
## [9]  {UVA PIRATA PETIT VERDOT 6U - 400253}      => {OSTRAS PEDRIN VERDOSILLA 6U - 300273}     0.01020
## [10] {OSTRAS PEDRIN VERDOSILLA 6U - 300273}     => {UVA PIRATA PETIT VERDOT 6U - 400253}      0.01020
## [11] {LANURIETA 75CL BLANCO 6U - 72322,
##       MARIA GAMBERRA 75CL TINTO 6U - 72314}     => {CAMI TINTO 75CL 6U - 73071}               0.01334
```

5.3 Basket Analysis Part2

This time with product category

```
ddi_table2 <-
  ddi %>% select(Store_name, Product_category) %>%
  group_by(Store_name, Product_category) %>%
  summarise(exists = TRUE, .groups = "drop") %>%
  pivot_wider(names_from = "Product_category", values_from = "exists") %>%
  replace(is.na(.), FALSE)
```

The result is a table with 3199 Stores & 9 Product Categories.

From the transaction table we get the transaction matrix doing:

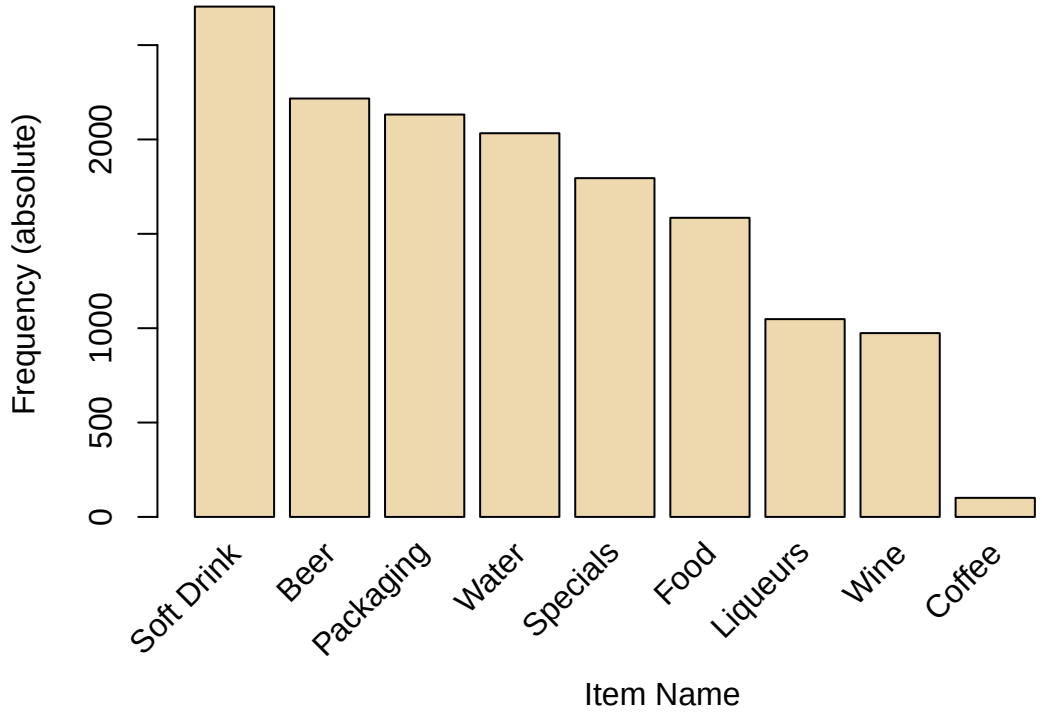
```
ddi_matrix2 <- as(ddi_table2 %>% select(-Store_name), "transactions")

summary(ddi_matrix2)
```

```
## transactions as itemMatrix in sparse format with
##  3199 rows (elements/itemsets/transactions) and
##  9 columns (items) and a density of 0.5067209
##
## most frequent items:
## Soft Drink       Beer  Packaging      Water   Specials    (Other)
##       2704       2217       2132       2033       1795       3708
##
## element (itemset/transaction) length distribution:
## sizes
##   1   2   3   4   5   6   7   8   9
## 417 464 344 330 379 401 373 439  52
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    5.00    4.56    7.00    9.00
##
## includes extended item information - examples:
##       labels  variables levels
## 1       Beer       Beer   TRUE
## 2  Packaging  Packaging   TRUE
## 3 Soft Drink Soft Drink   TRUE
##
## includes extended transaction information - examples:
##   transactionID
```

```
## 1              1
## 2              2
## 3              3
```

Right skewed distribution, checked

```
# Absolute Item Frequency Plot
itemFrequencyPlot(ddi_matrix2, topN = 10, type = "absolute", col = "wheat2", xlab = "Item Name", ylab =
```



have the highest frequency in the transactions.

5.3.1 Rules mining

Let's mine some rules on ddi_matrix (we'll have more rules if we lower minimal values of supp and conf). Since we have not many products like the previous part, this time we consider lower parameters to obtain more rules.

```
rules_ddi2 <- apriori(ddi_matrix2,
                 control = list(verbose = FALSE),
                 parameter = list(supp = 0.001, conf = 0.01))


summary(rules_ddi2)
```

```
## set of 2304 rules
```

```
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3   4   5   6   7   8   9
##   9  72 252 504 630 504 252  72   9
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       4       5       5       6       9
##
## summary of quality measures:
##     support          confidence         coverage            lift
##  Min.   :0.01626   Min.   :0.03157   Min.   :0.01626   Min.   :1.000
##  1st Qu.:0.01969   1st Qu.:0.75653   1st Qu.:0.02282   1st Qu.:1.383
##  Median :0.02657   Median :0.91394   Median :0.17162   Median :1.648
##  Mean   :0.11678   Mean   :0.79583   Mean   :0.17050   Mean   :1.810
##  3rd Qu.:0.20350   3rd Qu.:0.96667   3rd Qu.:0.25133   3rd Qu.:2.235
##  Max.   :0.84526   Max.   :1.00000   Max.   :1.00000   Max.   :3.557
##      count
##  Min.   :  52.0
##  1st Qu.:  63.0
##  Median :  85.0
##  Mean   : 373.6
##  3rd Qu.: 651.0
##  Max.   :2704.0
##
## mining info:
##         data ntransactions support confidence
##  ddi_matrix2          3199   0.001       0.01
```

Greater lift values indicate stronger associations so we sort the rules based on their lift.

```
#sort them by lift
rules_ddi2 <- sort(rules_ddi2, by="lift", decreasing = TRUE)
```

In this part we can find out how other products would have a basket relationship with Coffee.

```
rules_coffee <- subset(rules_ddi2, subset = rhs %in% "Coffee" & lift > 3.45)
inspect(rules_coffee)
```

```
##      lhs             rhs         support confidence  coverage     lift count
## [1]  {Beer,
##       Packaging,
##       Soft Drink,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01625508  0.1123110 0.1447327 3.557257    52
## [2]  {Beer,
##       Soft Drink,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
```

```
##       Wine}       => {Coffee} 0.01750547  0.1120000 0.1562988 3.547406    56
## [3]  {Soft Drink,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01781807  0.1115460 0.1597374 3.533026    57
## [4]  {Beer,
##       Packaging,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01625508  0.1111111 0.1462957 3.519252    52
## [5]  {Packaging,
##       Soft Drink,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01625508  0.1108742 0.1466083 3.511748    52
## [6]  {Beer,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01750547  0.1102362 0.1587996 3.491541    56
## [7]  {Beer,
##       Packaging,
##       Soft Drink,
##       Water,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01625508  0.1099366 0.1478587 3.482051    52
## [8]  {Packaging,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01625508  0.1097046 0.1481713 3.474704    52
## [9]  {Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01781807  0.1094050 0.1628634 3.465214    57
## [10] {Beer,
##       Soft Drink,
##       Water,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01750547  0.1093750 0.1600500 3.464264    56
## [11] {Beer,
##       Packaging,
##       Soft Drink,
```

```
##       Water,
##       Specials,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01688028  0.1093117 0.1544233 3.462260    54
## [12] {Soft Drink,
##       Water,
##       Food,
##       Liqueurs,
##       Wine}       => {Coffee} 0.01781807  0.1089866 0.1634886 3.451962    57
```

As the result show, the rules with highest lift have a support of less than 2 percent. This shows us that Coffee has lower transactions than the other product.

The same method can be implemented for Wine category.

```
rules_wine <- subset(rules_ddi2, subset = rhs %in% "Wine" & lift > 2.6)
inspect(rules_wine)
```

```
##      lhs             rhs       support confidence   coverage     lift count
## [1]  {Beer,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.8055556 0.02250703 2.645762    58
## [2]  {Beer,
##       Soft Drink,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.8055556 0.02250703 2.645762    58
## [3]  {Beer,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01875586  0.8000000 0.02344483 2.627515    60
## [4]  {Beer,
##       Soft Drink,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01875586  0.8000000 0.02344483 2.627515    60
## [5]  {Beer,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01750547  0.8000000 0.02188184 2.627515    56
## [6]  {Beer,
##       Soft Drink,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01750547  0.8000000 0.02188184 2.627515    56
## [7]  {Beer,
##       Water,
```

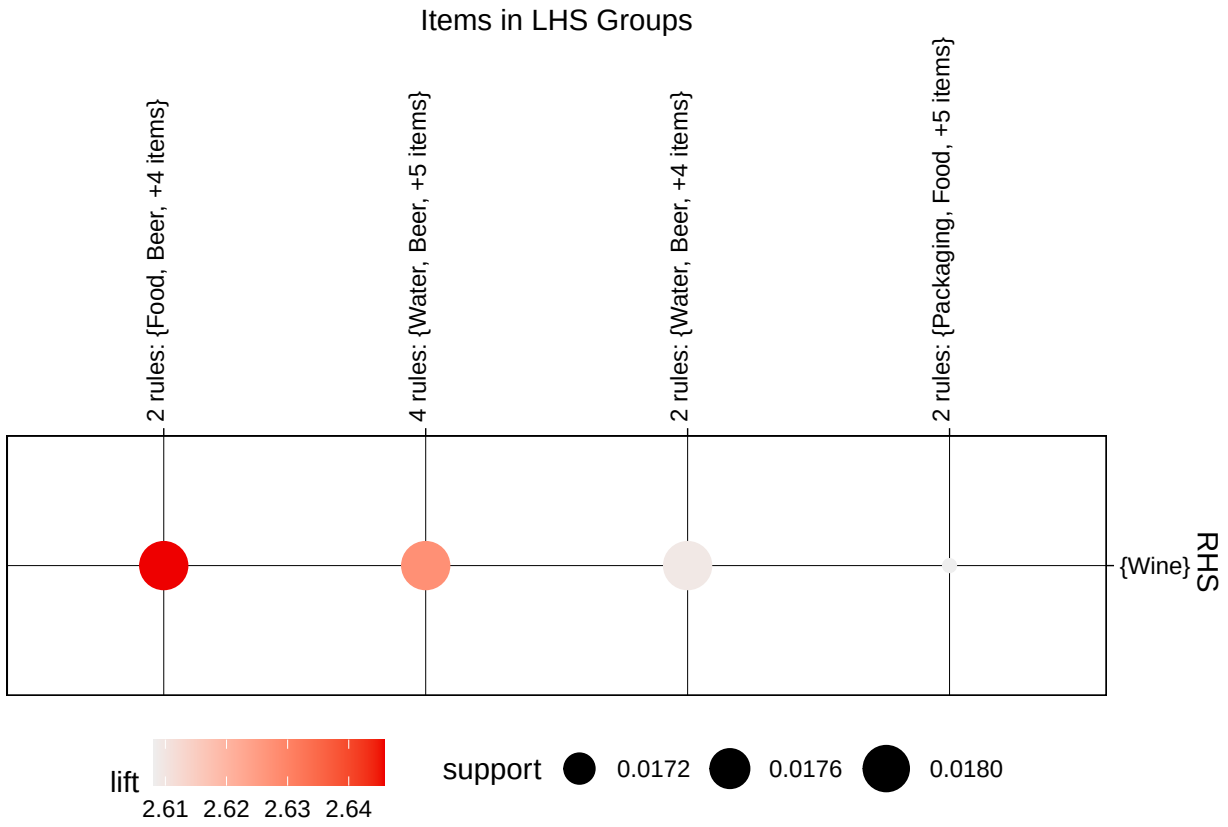
```
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.7945205 0.02281963 2.609519    58
## [8]  {Beer,
##       Soft Drink,
##       Water,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.7945205 0.02281963 2.609519    58
## [9]  {Beer,
##       Packaging,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01688028  0.7941176 0.02125664 2.608195    54
## [10] {Beer,
##       Packaging,
##       Soft Drink,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01688028  0.7941176 0.02125664 2.608195    54
```

```
rules_wine <- subset(rules_ddi2, subset = rhs %in% "Wine" & lift > 2.6)


rules_wine2 <- subset(rules_wine, subset = lhs %in% "Beer")
inspect(rules_wine2)
```

```
##      lhs             rhs       support confidence   coverage     lift count
## [1]  {Beer,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.8055556 0.02250703 2.645762    58
## [2]  {Beer,
##       Soft Drink,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.8055556 0.02250703 2.645762    58
## [3]  {Beer,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01875586  0.8000000 0.02344483 2.627515    60
## [4]  {Beer,
##       Soft Drink,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01875586  0.8000000 0.02344483 2.627515    60
## [5]  {Beer,
##       Water,
##       Specials,
##       Food,
```

```
##       Liqueurs,
##       Coffee}     => {Wine} 0.01750547  0.8000000 0.02188184 2.627515    56
## [6]  {Beer,
##       Soft Drink,
##       Water,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01750547  0.8000000 0.02188184 2.627515    56
## [7]  {Beer,
##       Water,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.7945205 0.02281963 2.609519    58
## [8]  {Beer,
##       Soft Drink,
##       Water,
##       Specials,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01813067  0.7945205 0.02281963 2.609519    58
## [9]  {Beer,
##       Packaging,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01688028  0.7941176 0.02125664 2.608195    54
## [10] {Beer,
##       Packaging,
##       Soft Drink,
##       Specials,
##       Food,
##       Liqueurs,
##       Coffee}     => {Wine} 0.01688028  0.7941176 0.02125664 2.608195    54
```

```
library(arulesViz)
# Graph (default layout)
plot(rules_wine, method="grouped")
```



This shows the correlation between the rules and also comparing their parameters.

6 Customer Segmentation using RFM Analysis

RFM is a method used for analyzing customer value. It is based on pareto principle: 20% of customers generate 80% of revenue.

RFM stands for the three dimensions:

Recency – How recently did the customer purchase? Frequency – How often do they purchase? *Monetary Value – How much do they spend?

RFM measures how much one customer contributes to a business. Recency tells when the most recent purchase of a particular customer was, frequency measures how frequently a customer buys from the business and monetary measures how much a customer spends at the business on average. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value). Identifying the most valuable RFM segments can capitalize on chance relationships in the data used for this analysis.

6.1 Data Cleaning

Prepare the data set. We need Store name, transaction date, Order quantity, Revenue.

```
#select columns we need
rfm_prep <- ddi %>% select(Store_name, contains(c("UMB","Euros")))
```

6.1.1 Step1: Calculate Number of Invoices

```
InvoiceNo <- ddi %>% select(Store_name) %>% group_by(Store_name) %>%
  summarise(NoInvoices=n())
```

6.1.2 Step2: Calculate Order Quantity

```
#build quantity data set
quantity <- ddi %>% select(Store_name, contains("UMB"))
#replace - with zero
quantity[quantity == "-"] <- "0"
#make columns numeric again
quantity[,-1] <- sapply(quantity[,-1],as.numeric)
#remove all-zero rows
quantity_clean <- quantity[rowSums(quantity[,-1])>0, ]
#get data for each store & month
quantity_grp <- quantity_clean %>%
  group_by(Store_name) %>%
  summarise(across(everything(), list(sum)))
#pivot table
quantity_pivot <- quantity_grp %>%
  pivot_longer(cols = contains("UMB"),
               names_to = "Date", values_to = "Quantity")
#remove zero rows
quantity_pivot_nozero <- quantity_pivot[rowSums(quantity_pivot[,3])>0, ]
```

```
quantity_pivot_nozero$Date <-
  revalue(quantity_pivot_nozero$Date,
                                      c("jun UMB  ant._1" = "2019/06/01",
                                        "jul UMB  ant._1" = "2019/07/01",
                                        "ago UMB  ant._1" = "2019/08/01",
                                        "sep UMB  ant._1" = "2019/09/01",
                                        "oct UMB  ant._1" = "2019/10/01",
                                        "nov UMB  ant._1" = "2019/11/01",
                                        "dic UMB  ant._1" = "2019/12/01",
                                        "ene UMB  ant._1" = "2019/01/01",
                                        "feb UMB  ant._1" = "2019/02/01",
                                        "mar UMB  ant._1" = "2019/03/01",
                                        "abr UMB  ant._1" = "2019/04/01",
                                        "may UMB  ant._1" = "2019/05/01",
                                        "jun UMB_1" = "2020/06/01",
                                        "jul UMB_1" = "2020/07/01",
                                        "ago UMB_1" = "2020/08/01",
                                        "sep UMB_1" = "2020/09/01",
                                        "oct UMB_1" = "2020/10/01",
                                        "nov UMB_1" = "2020/11/01",
                                        "dic UMB_1" = "2020/12/01",
                                        "ene UMB_1" = "2020/01/01",
                                        "feb UMB_1" = "2020/02/01",
                                        "mar UMB_1" = "2020/03/01",
```

```
                                        "abt UMB_1" = "2020/04/01",
                                        "may UMB_1" = "2020/05/01"))
```

```
## Warning: 'plyr' namespace cannot be unloaded:
##   namespace 'plyr' is imported by 'pROC' so cannot be unloaded
```

6.1.3 Step3: Calculate Revenue

```
#build revenue data set
revenue <- ddi %>% select(Store_name, contains("Euros"))
#replace - with zero
revenue[revenue == "-"] <- "0"
#make columns numeric again
revenue[,-1] <- sapply(revenue[,-1],as.numeric)
#remove all-zero rows
revenue_clean <- revenue[rowSums(revenue[,-1])>0, ]
#get data for each store & month
revenue_grp <- revenue_clean %>% group_by(Store_name) %>% summarise(across(everything(), list(sum)))
#pivot table
revenue_pivot <- revenue_grp %>% pivot_longer(cols = contains("Euros"), names_to = "Date", values_to = "
#remove zero rows
revenue_pivot_nozero <- revenue_pivot[rowSums(revenue_pivot[,3])>0, ]
```

```
revenue_pivot_nozero$Date <- revalue(revenue_pivot_nozero$Date,
                                     c("jun Euros  ant._1" = "2019/06/01",
                                       "jul Euros  ant._1" = "2019/07/01",
                                       "ago Euros  ant._1" = "2019/08/01",
                                       "sep Euros  ant._1" = "2019/09/01",
                                       "oct Euros  ant._1" = "2019/10/01",
                                       "nov Euros  ant._1" = "2019/11/01",
                                       "dic Euros  ant._1" = "2019/12/01",
                                       "ene Euros  ant._1" = "2019/01/01",
                                       "feb Euros  ant._1" = "2019/02/01",
                                       "mar Euros  ant._1" = "2019/03/01",
                                       "abr Euros  ant._1" = "2019/04/01",
                                       "may Euros  ant._1" = "2019/05/01",
                                       "jun Euros_1" = "2020/06/01",
                                       "jul Euros_1" = "2020/07/01",
                                       "ago Euros_1" = "2020/08/01",
                                       "sep Euros_1" = "2020/09/01",
                                       "oct Euros_1" = "2020/10/01",
                                       "nov Euros_1" = "2020/11/01",
                                       "dic Euros_1" = "2020/12/01",
                                       "ene Euros_1" = "2020/01/01",
                                       "feb Euros_1" = "2020/02/01",
                                       "mar Euros_1" = "2020/03/01",
                                       "abt Euros_1" = "2020/04/01",
                                       "may Euros_1" = "2020/05/01"))
```

```
## Warning: 'plyr' namespace cannot be unloaded:
##   namespace 'plyr' is imported by 'pROC' so cannot be unloaded
```

6.1.4 Step4: Recode Variables

We should do some recoding and convert character variables to factors.

```
quantity_rfm <- quantity_pivot_nozero %>%
  mutate(Store_name=as.factor(Store_name),
         Date=as.Date(Date, '%Y/%m/%d'))


revenue_rfm <- revenue_pivot_nozero %>%
  mutate(Store_name=as.factor(Store_name),
         Date=as.Date(Date, '%Y/%m/%d'))
```

```
rfm_prep <- merge(revenue_rfm, quantity_rfm, by = c("Store_name","Date"), all = FALSE)
rfm_prep <- merge(rfm_prep, InvoiceNo, by.x = "Store_name", all = FALSE)
rfm_prep <- rfm_prep %>% drop_na()
glimpse(rfm_prep)
```

```
## Rows: 43,454
## Columns: 5
## $ Store_name <fct> "(APARTAMENTOS APOLO) ESCALETA DEL MAR - 28979", "(APARTAME~
## $ Date       <date> 2019-07-01, 2020-06-01, 2019-08-01, 2019-09-01, 2019-10-01~
## $ revenue    <dbl> 421.92, 273.32, 529.00, 52.90, 52.90, 52.90, 228.00, 76.35,~
## $ Quantity   <dbl> 10, 7, 10, 1, 1, 1, 4, 2, 15, 14, 25, 16, 18, 14, 26, 30, 2~
## $ NoInvoices <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 7, 7, 7, 7, 7, 7, 7~
```

```
str(rfm_prep)
```

```
## 'data.frame':    43454 obs. of  5 variables:
##  $ Store_name: Factor w/ 3199 levels "(APARTAMENTOS APOLO) ESCALETA DEL MAR - 28979",..: 1 1 1 1 1 1
##  $ Date      : Date, format: "2019-07-01" "2020-06-01" ...
##  $ revenue   : num  421.9 273.3 529 52.9 52.9 ...
##  $ Quantity  : num  10 7 10 1 1 1 4 2 15 14 ...
##  $ NoInvoices: int  10 10 10 10 10 10 10 10 10 10 ...
```

```
kable(tail(rfm_prep))
```

	Store_name	Date	revenue	Quantity	NoInvoices
43449	ZURICH BAR - 30225	2020-07-01	1037.79	150	65
43450	ZURICH BAR - 30225	2020-08-01	1220.54	181	65
43451	ZURICH BAR - 30225	2020-09-01	676.75	86	65
43452	ZURICH BAR - 30225	2020-10-01	288.03	45	65
43453	ZURICH BAR - 30225	2020-11-01	184.29	31	65
43454	ZURICH BAR - 30225	2020-12-01	266.59	58	65

6.1.5 Step5: Calculate RFM

To implement the RFM analysis, we need to further process the data set in by the following steps:

1. Find the most recent date for each Store and calculate the days to the now (in this case we chose 01/06/2021) , to get the Recency data.

2. Calculate the quantity of transactions of a Store, to get the Frequency data.
3. Sum the amount of money a Store spent and divide it by Frequency, to get the amount per transaction on average, that is the Monetary data.

```
rfm_ddi <- rfm_prep %>%
  group_by(Store_name) %>%
  summarise(recency=as.numeric(as.Date("2021-06-01")- max(Date)),
            frequency=NoInvoices,
            monitery= sum(revenue)) %>%
  distinct(Store_name, .keep_all = TRUE)
```

```
## `summarise()` has grouped output by 'Store_name'. You can override using the `.groups` argument.
```

```
summary(rfm_ddi)
```

```
##                                          Store_name      recency
##  (APARTAMENTOS APOLO) ESCALETA DEL MAR - 28979:   1   Min.   :182.0
##  *ASOC CULTURAL CANNABICA - 32626              :   1   1st Qu.:182.0
##  *DOLCEMANIA (FONTDOR)- - 33562                :   1   Median :182.0
##  *F.DE FISICA I QUI (FONTDOR) - 33520          :   1   Mean   :338.4
##  *FERROVIAL (FONTDOR) BCN SANTS - 32093        :   1   3rd Qu.:457.0
##  *GRUPO LA RAUNION (FONTDOR) - 31522           :   1   Max.   :882.0
##  (Other)                                       :3097
##    frequency         monitery
##  Min.   :  1.00   Min.   :     0.23
##  1st Qu.: 10.00   1st Qu.:  1401.04
##  Median : 20.00   Median :  5225.37
##  Mean   : 30.39   Mean   :  9052.98
##  3rd Qu.: 40.00   3rd Qu.: 12942.10
##  Max.   :370.00   Max.   :150490.69
##
```

```
kable(tail(rfm_ddi))
```

Store_name	recency	frequency	monitery
ZIM BAR - 23482	457	3	1792.99
ZIRYAB SHISHA LOUNGE - 24037	701	6	104.31
ZITAROSA - 27858	182	78	65867.29
ZUM ZEIG BISTROT - 31723	486	12	931.51
ZURICH BAR - 09772	182	16	11998.44
ZURICH BAR - 30225	182	65	11649.29

6.2 RFM Analysis

We will calculate the scores based on the quartiles obtained from the summary.

```
#Scoring
#R_score
rfm_ddi$R_Score[rfm_ddi$recency>457]<-1
```

```
## Warning: Unknown or uninitialised column: `R_Score`.
```

```
rfm_ddi$R_Score[rfm_ddi$recency>280 & rfm_ddi$recency<=457 ]<-2
rfm_ddi$R_Score[rfm_ddi$recency>220 & rfm_ddi$recency<=280]<-3
rfm_ddi$R_Score[rfm_ddi$recency<=220]<-4

#F_score
rfm_ddi$F_Score[rfm_ddi$frequency<10]<-1
```

```
## Warning: Unknown or uninitialised column: `F_Score`.
```

```
rfm_ddi$F_Score[rfm_ddi$frequency>=10 & rfm_ddi$frequency<20]<-2
rfm_ddi$F_Score[rfm_ddi$frequency>=20 & rfm_ddi$frequency<40 ]<-3
rfm_ddi$F_Score[rfm_ddi$frequency>=40]<-4

#M_score
rfm_ddi$M_Score[rfm_ddi$monitery<= 12942.10]<-1
```

```
## Warning: Unknown or uninitialised column: `M_Score`.
```

```
rfm_ddi$M_Score[rfm_ddi$monitery>=12942.10 & rfm_ddi$monitery<5225.37]<-2
rfm_ddi$M_Score[rfm_ddi$monitery>=5225.37 & rfm_ddi$monitery<1401.04 ]<-3
rfm_ddi$M_Score[rfm_ddi$monitery>=1401.04]<-4

#RFM_score
rfm_ddi<- rfm_ddi %>% mutate(RFM_Score = 100*R_Score + 10*F_Score+M_Score)
```

6.3 Customer Segments

In this part the customers will be classified based on the recency, frequency and monetary scores.

```
#Customer Segmentation
rfm_ddi$segmentRFM <- NULL
champions <- c(444)
loyal_customers <- c(334, 342, 343, 344, 433, 434, 443)
potential_loyalist <- c(332,333,341,412,413,414,431,432,441,442,421,422,423,424)
recent_customers <- c(411)
promising <- c(311, 312, 313, 331)
needing_attention <- c(212,213,214,231,232,233,241,314,321,322,323,324)
at_risk <- c(112,113,114,211,141,131,132,133,142,124,123,122,121,224,223,222,221)
cant_lose <- c(134,143,144,234,242,243,244)
lost <- c(111)

rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% champions)] = "Champions"
```

```
## Warning: Unknown or uninitialised column: `segmentRFM`.
```

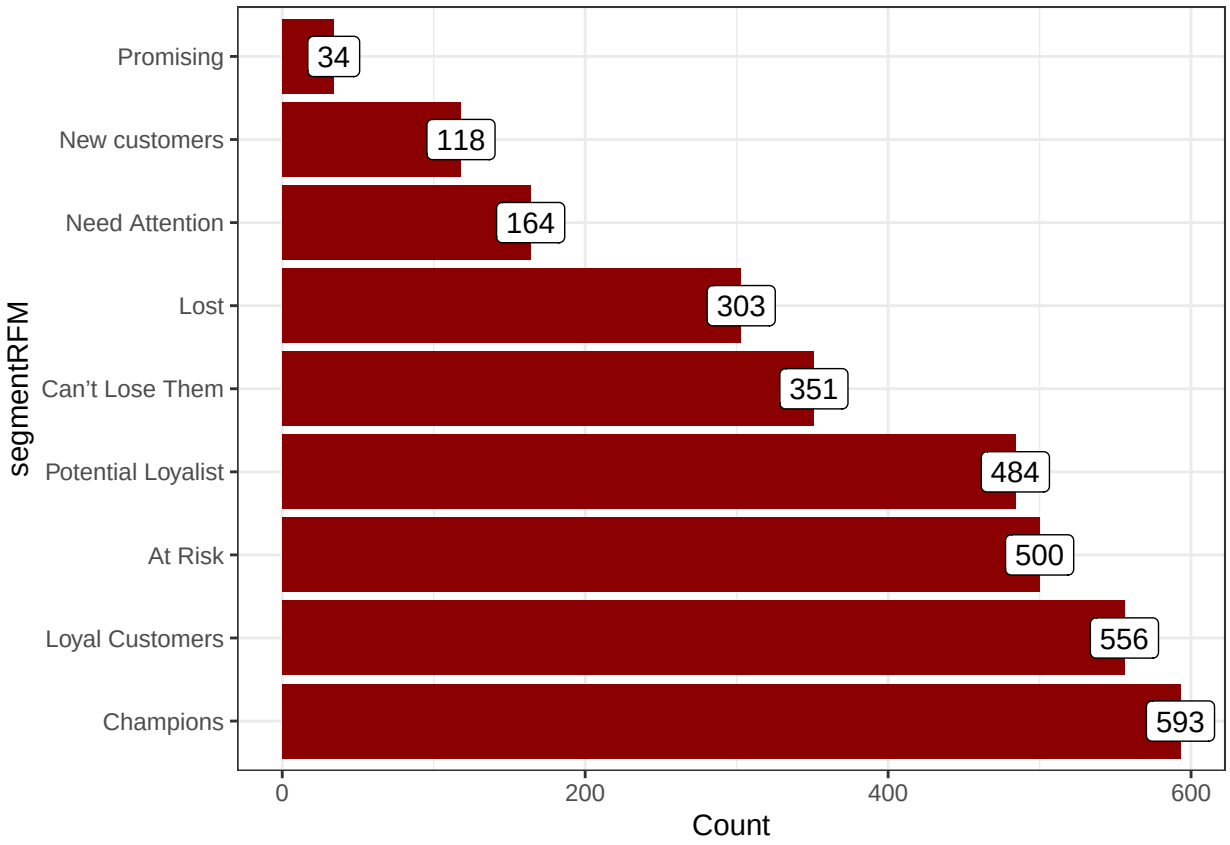
```
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% loyal_customers)] = "Loyal Customers"
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% potential_loyalist)] = "Potential Loyalist"
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% recent_customers)] = "New customers"
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% promising)] = "Promising"
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% needing_attention)] = "Need Attention"
```

```
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% at_risk)] = "At Risk"
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% cant_lose)] = "Can't Lose Them"
rfm_ddi$segmentRFM[which(rfm_ddi$RFM_Score %in% lost)] = "Lost"
```

Evaluate the segment size. Once we have classified a customer into a particular segment, we can take appropriate action to increase their lifetime value.

```
segment_size <-
 rfm_ddi %>%
  group_by(segmentRFM) %>%
 count(segmentRFM) %>%
 arrange(desc(n)) %>%
 rename(segmentRFM = segmentRFM, Count = n)
```

```
segment_size$segmentRFM <- factor(segment_size$segmentRFM,
levels = segment_size$segmentRFM[order(-segment_size$Count)])

segment_size %>%
  ggplot(aes(x=segmentRFM, y=Count)) +
  geom_bar(stat="identity",
           show.legend=FALSE, fill="darkred") +
  scale_color_brewer(palette = "BuPu") +
  geom_label(aes(label=Count)) +
  coord_flip() +
  theme_bw()
```



6.4 Recommendation

In conclusion, we have 9 segments in our dataset.

• 593 Champions are our best customers, who bought most recently, most often, and are heavy spenders. They can become early adopters for new products and will help promote our brand. • 484 Potential Loyalists are our recent customers with average frequency and who spent a good amount. Offer membership or loyalty programs or recommend related products to upsell them and help them become our Loyalists or Champions. • 500 At Risk Customers are our customers who purchased often and spent big amounts, but haven't purchased recently. Send them personalized reactivation campaigns to reconnect, and offer renewals and helpful products to encourage another purchase. • 351 Can't Lose Them are customers who used to visit and purchase quite often, but haven't been visiting recently. Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor. • 118 New Customers are our customers who have a high overall RFM score but are not frequent shoppers. Start building relationships with these customers by providing onboarding support and special offers to increase their visits.

Given the cost to create and launch a new product, combined with the rate of failure, being able to successfully develop and bring new products to market is an important capability.

7 Conclusion

This type of analysis can be conducted on the data set as the transactions get higher to understand the customer needs and leverage other products to introduce new ones. The market has a lot of potential for Coffee and Wine, thus, with the right strategy and using the historical data the company can escalate faster comparing to rivals.